IEEE Signal Processing MAGAZINE

Volume 35 | Number 5 | September 2018

THE INTERNET OF THINGS

Main Advances and Key Topics of Development

What Is a Signal?

Low-Rank Matrix Completion

Signal Processing Opens the IoT to New Possibilities

Highlights from the SP Cup 2018 Student Competition

IEEE Signal Processing Society

IEEE

Digital Object Identifier 10.1109/MSP.2018.2857438

Contents

Volume 35 | Number 5 | September 2018

SPECIAL SECTION

SIGNAL PROCESSING AND THE INTERNET OF THINGS

PG. 9

ON THE COVER

This special issue of *IEEE Signal Processing Magazine* provides a comprehensive view of the main advances in the Internet of Things through a number of tutorial-style articles as well as from contributions that emphasize the key topics of development in this area, both in terms of theory and applications.

COVER IMAGE: ©ISTOCKPHOTO.COM/ILYAST,
ROBOT, VEHICLE, UAV ICONS COURTESY OF NURIA GONZALEZ PRELCIC

PG. 168

IEEE SIGNAL PROCESSING MAGAZINE (ISSN 1053-5888) (ISPREG) is published bimonthly by the Institute of Electrical and Electronics Engineers, Inc., 3 Park Avenue, 17th Floor, New York, NY 10016-5997 USA (+1 212 419 7900). Responsibility for the contents rests upon the authors and not the IEEE, the Society, or its members. Annual member subscriptions included in Society fee. Nonmember subscriptions available upon request. **Individual copies:** IEEE Members US$20.00 (first copy only), nonmembers US$241.00 per copy. Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limits of U.S. Copyright Law for private use of patrons: 1) those post-1977 articles that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923 USA; 2) pre-1978 articles without fee. Instructors are permitted to photocopy isolated articles for noncommercial classroom use without fee. **For all other copying, reprint, or republication permission,** write to IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854 USA. Copyright © 2018 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Periodicals postage paid at New York, NY, and at additional mailing offices. **Postmaster:** Send address changes to IEEE Signal Processing Magazine, IEEE, 445 Hoes Lane, Piscataway, NJ 08854 USA. Canadian GST #125634188 **Printed in the U.S.A.**

Digital Object Identifier 10.1109/MSP.2018.2849040

COLUMNS

©ISTOCKPHOTO.COM/GIVAGA

The IEEE International Symposium on Biomedical Imaging will be held 8–11 April 2019 in beautiful Venice, Italy.

DEPARTMENTS

IEEE prohibits discrimination, harassment, and bullying. For more information, visit http://www.ieee.org/web/aboutus/whatis/policies/p9-26.html.

IEEE Signal Processing Magazine

Digital Object Identifier 10.1109/MSP.2018.2849041

SCOPE: *IEEE Signal Processing Magazine* publishes tutorial-style articles on signal processing research and applications as well as columns and forums on issues of interest. Its coverage ranges from fundamental principles to practical implementation, reflecting the multidimensional facets of interests and concerns of the community. Its mission is to bring up-to-date, emerging and active technical developments, issues, and events to the research, educational, and professional communities. It is also the main Society communication platform addressing important issues concerning all members.

Robert W. Heath, Jr. | Editor-in-Chief | rheath@utexas.edu

GlobalSIP and Beyond

Like many readers of *IEEE Signal Processing Magazine* (*SPM*), I have been involved with many activities in the IEEE Signal Processing Society (SPS). One of my favorite activities (besides, of course, attending) has been in organizing conferences. I was on the organizing team in a variety of capacities for SPS conferences including the IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), the International Workshop on Computational Advances in Multisensor Adaptive Processing, and the IEEE Global Conference on Signal and Information Processing (GlobalSIP) as well as conferences in other Societies including the IEEE International Symposium on Information Theory, IEEE Global Communications Conference (GLOBECOM), IEEE Vehicular Technology Conference, and IEEE Communication Theory Workshop, not to mention the Asilomar Conference on Signals, Systems, and Computers. Conferences are a great way to learn about new ideas, develop important collaborations, and, of course, explore the social spectrum of eating, dancing, and karaoke. Along these lines, I would like to devote this editorial to the GlobalSIP conference. This is timely as the SPS Board of Governors recently voted on a motion to conclude GlobalSIP after 2019 (for full disclosure, I was one of two opposing votes on this motion).

GlobalSIP was created to serve as a flagship conference for SPS in 2012. I was a founding general cochair with SPS's President-Elect Ahmed Tewfik. I would like to provide my perspective on the key features of the first GlobalSIP conference and also speculate on why the GlobalSIP series was not successful. These are my own opinions and observations.

There is so much value in special issues—I hope we can find a way to have such a feature in our conferences.

GlobalSIP was created to be a collection of colocated symposia, with a common daily keynote and social activity. These symposia were meant to be associated around emerging topics and not necessarily aligned with a single technical committee as in the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). In my mind, a main benefit of this approach was to pull in papers based on a common theme. At ICASSP, papers are binned into sessions after acceptance and sometimes the sessions are quite heterogeneous. I thought the idea with GlobalSIP was to organize symposia in the same way that we organize special issues for *SPM* or *IEEE Journal of Selected Topics in Signal Processing*: inviting organizers to submit a symposia proposal, advertising call for papers of selected symposia, and then hosting the event at the conference. It works for the magazine, and I thought it would work for GlobalSIP as well. Similar concepts are used at other conferences like the IEEE International Conference on Communications and GLOBECOM (called *workshops*) and are often well attended.

The GlobalSIP symposia were envisioned to each follow a similar template, with two sessions of posters and one of plenary talks. We chose posters because they are an excellent way to transfer information for those skilled in the area, and they are interactive. I enjoy the posters at ICASSP and SPAWC. We envisioned one flexible session where the organizers could invite a plenary speaker or organize a panel; most opted for a plenary talk. The overall structure seemed to be good based on the formats of other conferences.

There was one significant GlobalSIP organizational quirk. When having multiple colocated symposia, it did not make sense to have a plenary from every symposia at the same time. This also relaxed some room scheduling requirements. As a result, some symposia had their plenary talk in the early or late afternoon. In my opinion, the benefits were huge. Students or members in general who wanted to learn about new trends could attend plenaries all day! That said, it did create grumbling among the symposia organizers.

So why did GlobalSIP fail? For many people, it was "yet another conference." The timing was not ideal, occurring at the

(continued on page 15)

Digital Object Identifier 10.1109/MSP.2018.2849042
Date of publication: 28 August 2018

Ali H. Sayed | IEEE Signal Processing Society President | ali.sayed@epfl.ch

Science Is Blind

I started drafting this editorial on July 4th while sitting in my hotel room in Versailles, France. Both the date and location have great significance in our modern history, which motivated my choice for the theme of the article.

The date of July 4th coincides with the commemoration of Independence Day in the United States. It refers to the day back in 1776 when the Declaration of Independence, drafted by Thomas Jefferson and his colleagues, was adopted. The location, next to the Palace of Versailles, which housed the Kings of France until the outbreak of the French Revolution in 1789, reminded me of a second historical document approved that year by the French Constituent Assembly and known as the *Declaration of the Rights of Man and of the Citizen*. This document was also drafted with input from Thomas Jefferson. Both documents ascertained the rights of men and served as drivers for civil liberties, although with some challenges along the way.

Today, we experience a continuous stream of news, stereotypes, and opinions about immigrants and foreigners, including veiled arguments hinting at the superiority of one race or ethnicity over another. As scientists, we value diversity in all its forms and know that science and education should help reduce inequities due to racial, ethnic, gender, religious, or economic biases.

I am the son of immigrants. My parents immigrated in the 1950s to the faraway, beautiful, and generous land of Brazil where I was born. Later in life, I followed in their footsteps and immigrated to the United States, the most creative and inventive land on Earth, a land of opportunities, one that was described as the "shining city on the hill" welcoming hard-working people from all corners of Earth with its majestic Statue of Liberty. The statue itself was a gift from the people of France to the American people in 1886; a second historical link between the two countries besides the declarations mentioned previously. One of the main reasons for the prominence of the United States on the World stage today is that it embraced diversity, pushed for equality, and opened its doors to the best and brightest who helped propel a wave of innovation and economic growth.

In this article, we focus on science and discovery, and on how diversity in all forms empowers both of them.

The Founding Fathers wrote in the U.S. Declaration of Independence that "We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty, and the pursuit of Happiness." I have always marveled at the beauty, simplicity, and clarity of this statement. Yet, this sentence has had its sufficient share of criticism since its early days. Some have wondered how a statement that "declares all men to be equal" could coexist with segregation and slavery. Abraham Lincoln responded to this criticism by arguing that the statement represented an ideal that the country should be striving to achieve. Others argued that the statement should have declared that "all men *and* women" are equal as a stepping stone toward gender equality. And yet others question whether "all men" refer to citizens only or should include other residents as well. The title of the French declaration mentions both "men" and "citizens" and went on to state similarly that "men are born and remain free and equal in rights." It is amazing how even in these honest attempts at declaring equality, the language can sometimes fail us. Today we understand the statements in a broad figurative sense in that they are blind to gender, race, ethnicity, origin, or religion. There is no question that the United States, with its generous embrace of diversity and immigrants, has led the way in pushing the frontiers of the human pursuit of opportunity, creativity, and ingenuity.

In this article, we focus on science and discovery, and on how diversity in all forms empowers both of them. Science should be blind to borders or national origin, race or gender, ethnicity

Digital Object Identifier 10.1109/MSP.2018.2849043
Date of publication: 28 August 2018

or religion, or any other alien consideration. Curiosity, ideas, and ideals are innate to the human condition; they cannot be confined to particular groups or boundaries and have wings of their own. The statistics speak loudly in support of diversity and inclusiveness and show how they have been drivers of innovation in the United States and other developed countries. Roughly one-third of all Nobel Prizes in the physics, chemistry, medicine, and economics fields received by Americans have been awarded to foreign-born scientists [1]. And according to a 2018 report [2] by the U.S. National Science Foundation, foreign-born individuals account for about 30% of college-educated workers in the United States in science and engineering. Among workers with Ph.D. degrees, the percentage is higher at 42%. Similar figures apply to countries in Europe. According to a 2012 working paper from the National Bureau of Economic Research [3], approximately 57% of scientists in Switzerland are foreign born (with Germany being the main feeder), 38% of scientists in Sweden are foreign born, 33% in the United Kingdom, 28% in The Netherlands, 23% in Germany, 22% in Denmark, and 17% in France. That figure is 47% in Canada and 45% in Australia. These numbers were based on surveying over 17,000 scientists in 2011.

It is not difficult to name several famous immigrant scientists who have revolutionized science and technology in the United States such as Nikola Tesla (originally from Serbia) and the Nobel Laureates Niels Bohr (Denmark), Albert Einstein (Germany), Enrico Fermi (Italy), and Ahmed Zewail (Egypt). We can also list entrepreneurs such as Sergey Brin (originally from Russia, cofounder of Google), Jerry Yang (Taiwan, cofounder of Yahoo!), Amar Bose (India, founder of Bose), as well as the cofounders of YouTube Steve Chen (Taiwan) and Jawed Karim (born in Germany to a Bangladeshi father). Steve Jobs himself (cofounder of Apple) was the son of a Syrian immigrant. Jeff Bezos (founder of Amazon) was adopted by a Cuban immigrant. If we also examine the list of recipients of the U.S. National Medal of Science in the domain of engineering sciences on Wikipedia we will find an extensive list of foreign-born awardees. I counted approximately 22 between 1962 and 2012 including, from our discipline, names like A. Viterbi (2007, Italian-American), R. Kalman (2008, Hungarian-American), and T. Kailath (2012, Indian-American). To my surprise, the list contains hardly any female recipients.

Yet the significant contributions of female engineers and scientists should not go unnoticed. I enjoyed watching the 2016 movie *Hidden Figures* with my daughters. It tells the wonderful story of a group of African-American female mathematicians working for NASA during the early development of the U.S. space program in the late 1950s. These mathematicians were referred to as *computers* within NASA. They were placed in segregated offices and had to use separate restroom facilities. Nevertheless, they persevered and earned the respect of their colleagues with superb grace, determination, and utter qualification. One of them, named Katherine Johnson, was entrusted with checking the trajectory for John Glenn's spacecraft. She received the U.S. Presidential Medal of Freedom in 2015 from President Barack Obama. These women surpassed racial and gender biases during their time and left a lasting mark on the history of the U.S. space program.

> **As an international scientific organization, we have the duty to project a model of inclusiveness given our diverse membership which is spread over all continents.**

Science prides itself of objectivity. Scientists and academicians tend to view themselves as unbiased individuals upholding the highest standards of fairness. They presume that they are color- or race- or gender-blind. Unfortunately, this is not always the case even in modern times, and we need to remain vigilant. Many known examples exist. Consider the experience of Emmy Noether (1882–1935), who is considered to be one of the greatest mathematical minds of the 20th century, and yet she is unknown to most of us [4]. She was a pioneer in the field of abstract algebra. She was not able to secure a professorship in Germany due to her gender and ethnicity, and had to teach her courses at the University of Gottingen under the name of another male colleague, none other than David Hilbert (of Hilbert space fame). Even Marie Curie (1867–1934), the two-time winner of the Nobel Prize in Physics (1903) and Chemistry (1911), and who is an inspiration today to women worldwide in the STEM fields, had her application to join the French Academy of Sciences rejected [5]! Marie Curie was also a foreign-born scientist: born in Poland and naturalized French. Moving closer to statistical signal processing, consider the story of David Blackwell (1919–2010), of Rao-Blackwell Theorem fame in mathematical statistics and a student of J.L. Doob. He made superlative contributions to Bayesian statistics, dynamic programming, and game theory. He had to leave his postdoctoral position at the Institute of Advanced Studies due to objections about his race at Princeton University in the early 1940s [6]. He ended up being the first black faculty member to join the University of California at Berkeley in 1955, and the first black American inducted into the U.S. National Academy of Sciences in 1965.

Scientists are not immune to racism, even someone as notable as the 1921 Nobel Laureate Albert Einstein (1879–1955). Many of us were startled to read this past June about entries in his 1922–1923 travel diaries revealing appalling remarks about "Chinese" and "Levantines of every shade." The comments were written when Einstein was in his 40s and still living in Europe. It is conflicting to believe that this is the same person who, after moving to the United States in 1933, spoke against racial segregation. Science itself is not immune to racism either, and has been used in the past, and even today, to advance prejudices and to justify the superiority of one race over another.

A moral imperative for scientific organizations

As an international scientific organization, we have the duty to project a model of inclusiveness given our diverse membership which is spread over all continents. While 40% of our members in the IEEE Signal Processing Society originate from the United States, 27% of them are from Asia; another 27% are from Europe, Africa, and the Middle East; 3% are from Canada; and 3% are from Central and South America. The fact that the numbers are aggregated for Europe, Africa, and the Middle East is an anomaly of the IEEE accounting system, which should be fixed. The practice of aggregating statistics is misleading because it masks challenges that may exist in certain regions.

We also continue to be dominated by a largely male membership base accounting for 80% of our members. We are working tirelessly to enlarge our pool of female members, which is part of a broader effort toward attracting more women to the STEM fields where they continue to be a minority. Let us not forget that even in the United States, women were not admitted into the undergraduate programs of many Ivy League schools (including Harvard, Yale, and Princeton) until the late 1960s and early 1970s. Also, in the United States, only about 25% of STEM graduates are women; the figure drops to 17% for the electrical engineering discipline. A recent study [7] suggests that regions in the world where the gender gap is smaller (such as countries in Europe or the United States) tend to have a smaller percentage of women in the STEM fields than regions where the gender gap is larger (such as some countries in Africa or the Middle East). For example, 41% of STEM members are women in Algeria, while that figure drops to less than 20% in The Netherlands and Belgium. This observation speaks in favor of diversity: one way toward increasing the representation of the STEM fields may be to have more openness toward regions with larger gender gap (where salaries and opportunities for women are more limited).

In another effort by our Society toward increased diversity, we are reaching out to students from all backgrounds, especially students from underprivileged regions.

In another effort by our Society toward increased diversity, we are reaching out to students from all backgrounds, especially students from underprivileged regions. We have adopted a US$1 per year membership policy for all students. We are now the US$1 student Society within the IEEE; so tell your students and tell your friends. Our Society has also affirmed its commitment to be considerate of the diversity of its members in all its activities, including publications and conferences. We experience diversity in many ways. We experience it in every conference we attend, with literally hundreds or thousands of attendees flying in from different regions. We experience diversity in every paper we read with authors from diverse countries, and in every lecture we give with curious faces from varied backgrounds looking at us with eagerness to learn and understand. We are an open Society. We understand that, given an opportunity, each member can make a contribution and have an impact. We have no borders in our professional Society. Your science pushes you forward. Our activities serve as a melting pot where cultures converge; scientists of different races and ethnicities; of different cultural backgrounds and religions, all standing equal under science. Our diversity, life, and work experiences enrich our scientific debates.

We live in a global world today where we are constantly reminded that the human condition has innate biases and suspicions in it. Yet, as scientists, we should keep an open mind and use science to promote understanding and inclusiveness. Once, when moving through an immigration line at a U.S. airport, an immigration officer requested to see my passport. Looking at me, she asked respectfully: “Mr. Sayed, how come you have a Middle-Eastern name but were born in Brazil?” Sensing that she was approachable, I looked at the name tag on her shirt and responded: “Ma’am, just like you, you have a beautiful Asian name and speak perfect English.”

References

[1] J. Bruner. (2011, Oct. 5). American leadership in science, measured in Nobel prizes. *Forbes Magazine*. [Online]. Available: https://www.forbes.com/sites/jonbruner/2011/10/05/nobel-prizes-and-american-leadership-in-science-infographic

[2] U.S. National Science Foundation. (2018). Science and engineering indicators. [Online]. Available: https://www.nsf.gov/statistics/2018/nsb20181/

[3] C. Franzoni, G. Scellato, and P. Stephan. (2012, May). Foreign born scientists: Mobility patterns for sixteen countries. U.S. Nat. Bureau Economic Research. Cambridge, MA. Tech. Rep. 18067. [Online]. Available: http://www.nber.org/papers/w18067

[4] N. Angrier. (2012, Mar. 26). The mighty mathematician you’ve never heard off. *New York Times*. [Online]. Available: https://www.nytimes.com/2012/03/27/science/emmy-noether-the-most-significant-mathematician-youve-never-heard-of.html

[5] T. Long. (2011, Jan. 23). Science academy tells Marie Curie, Non. Wired. [Online]. Available: https://www.wired.com/2012/01/jan-23-1911-marie-curie/

[6] W. Grimes. (2010, July 17). David Blackwell, scholar of probability, dies at 91. *New York Times*. [Online]. Available: https://www.nytimes.com/2010/07/17/education/17blackwell.html

[7] O. Khazan. (2018, Feb. 18). The more gender equality, the fewer women in STEM. *The Atlantic*. [Online]. Available: https://www.theatlantic.com/science/archive/2018/02/the-more-gender-equality-the-fewer-women-in-stem/553592/

SP

Election of Regional Directors-at-Large and Members-at-Large

Your vote is important! The election is now open for regional directors-at-large for Regions 1–6 and 8 (the term is 1 January 2019 through 31 December 2020) and members-at-large (term 1 January 2019 through 31 December 2021) of the IEEE Signal Processing Society (SPS) Board of Governors (BoG). Ballots, which have been mailed to SPS members, include a diverse slate of candidates for both elections, which were vetted by the SPS Nominations and Appointments Committee, as well as a space for write-in candidates. This year's election offers SPS members the opportunity to cast their votes via the web at https://eballot4.votenet.com/IEEE for up to one regional director-at-large for your corresponding Region, Regions 1–6 (United States) and Region 8 (Europe, Africa, and Middle East), and three member-at-large candidates. Ballots must be received at the IEEE no later than 1 October 2018 to be counted. Members must meet the eligibility requirements at the time the ballot data are generated to be eligible to vote. To be eligible to vote in this year's Society election, you had to have been an active SPS member or affiliate (excluding student member) prior to 1 August 2018. This is the date when the list of eligible Society voting members was compiled. The candidates for regional director-at-large are:

- *Regions 1–6*
  - Iole Moccagatta
  - Bhuvana Ramabhadran
- *Region 8*
  - Cédric Richard
  - Raed Shubair.

The candidates for member-at-large are

- Magdy A. Bayoumi
- Paulo S.R. Diniz
- Eric Fosler-Lussier
- Hamid Krim
- Douglas O'Shaughnessy

Digital Object Identifier 10.1109/MSP.2018.2861478
Date of publication: 28 August 2018

The candidates for regional director-at-large

Regions 1–6

Iole Moccagatta

Bhuvana Ramabhadran

Region 8

Cédric Richard

Raed Shubair

The candidates for member-at-large

Magdy A. Bayoumi

Paulo S.R. Diniz

Eric Fosler-Lussier

Hamid Krim

Douglas O'Shaughnessy

Ana Isabel Pérez-Neira

Wan-Chi Siu

Zhi (Gerry) Tian

- Ana Isabel Pérez-Neira
- Wan-Chi Siu
- Zhi (Gerry) Tian.

The BoG is the governing body that oversees the activities of the SPS. The SPS BoG has the responsibility of establishing and implementing policy and receiving reports from its standing boards and committees and comprises 21 Society members: six officers of the Society who are elected by the BoG, nine members-at-large elected by the voting members of the Society, four regional directors-at-large elected locally by Society voting members of the corresponding region, as well as the Awards Board chair. The six officers are the president, president-elect, the vice president-conferences, vice president-membership, vice president-publications, and vice president-technical directions. The executive director of the Society shall serve ex-officio, without vote.

Regional directors-at-large are SPS members who are elected locally by Society voting members of the corresponding Region via the annual election to serve on the Society's BoG as nonvoting members and voting members of the Society's Membership Board.

Members-at-large represent the member viewpoint in the Board decision making. They typically review, discuss, and act upon a wide range of items affecting the actions, activities, and health of the Society. More information on the IEEE SPS can be found at http://www.signalprocessingsociety.org/.

New Society Officer Elected and Editors-in-Chief Named for 2019

The Board of Governors of the IEEE Signal Processing Society (SPS) elected one new officer who will start her term on 1 January 2019. In addition, three volunteers have been named as editors-in-chief of IEEE SPS publications. The term for these editors-in-chief will run from 1 January 2019 through 31 December 2021.

New Society officer elected

Tülay Adalı will serve as the 2019–2021 vice president-technical directions. She is a Fellow of the IEEE and is with the University of Maryland, Baltimore County. She succeeds Walter Kellermann, who has held this position since January 2016.

Incoming editors-in-chief

Lina Karam has been named editor-in-chief of *IEEE Journal of Selected Topics in Signal Processing*. She is a Fellow of the IEEE and is with Arizona State University. She is succeeding Shrikanth (Shri) S. Narayanan, University of Southern California, who has been the editor-in-chief since 2016.

Antonio Ortega is the new editor-in-chief of *IEEE Transactions on Signal and Information Processing over Networks*. He is a Fellow of the IEEE and is with the University of Southern California. He is succeeding Petar Djurić, State University of New York, who has been editor-in-chief since 2015.

Dilek Hakkani-Tur has taken on the role of editor-in-chief of *IEEE Transactions on Audio, Speech, and Language Processing*. She is a Fellow of the IEEE and is with the Amazon Alexa artificial intelligence team. She is succeeding Haizhou Li, National University of Singapore, who has held the position since 2015.

SP

Digital Object Identifier 10.1109/MSP.2018.2853818
Date of publication: 28 August 2018

John Edwards

Signal Processing Opens the Internet of Things to a New World of Possibilities

Research leads to new Internet of Things technologies and applications

The *Internet of Things* (*IoT*) refers to the wireless connection of ordinary objects, such as vehicles, cash machines, door locks, cameras, industrial controls, and municipal traffic systems, to the Internet. Research firm BI Intelligence predicts that 22.5 billion devices will be connected to the IoT in 2021, compared to 6.6 billion in 2016.

Signal processing is playing a significant role in expanding the number of IoT technologies and applications. Realizing that the IoT has emerged as perhaps the most important new technology since the arrival of the Internet itself, researchers worldwide are now turning to signal processing to support and augment new IoT services and make existing applications less expensive and more practical.

Up in the air

One of the most promising IoT applications is the use of small aerial drones to read radio-frequency identification (RFID) tags affixed to boxes, crates, and other objects inside distribution centers and similar storage sites. Looking to enhance the process, researchers from the Massachusetts Institute of Technology (MIT) have developed a system that allows small drones to read RFID tags from tens of meters away while identifying the tags' locations with an average error measured in centimeters. The researchers envision a system that could be used in large facilities for continuous monitoring to prevent inventory mismatches and pinpoint the location of individual items so employees can rapidly and reliably meet customer requests.

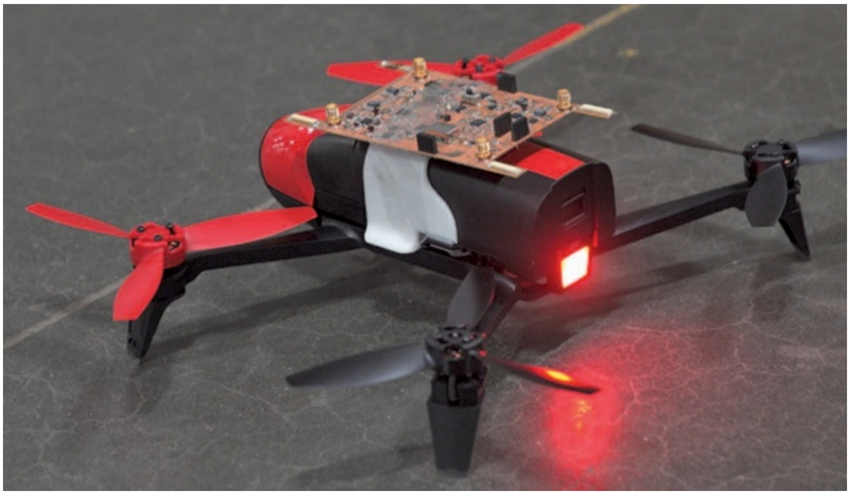

MIT

FIGURE 1. The RFly drone and relay circuit.

Wireless, battery-free RFID tags are inexpensive—approximately US$.05–.10 each—and disposable, offering a huge potential to revolutionize the logistics and supply-chain industry. "Today, their communication range is fundamentally crippled by their battery-free nature, demanding a reader standing nearby to perform wireless charging," explains lead researcher Yunfei Ma, a postdoctoral fellow in MIT's Signal Kinetics research group, headed by Fadel Adib, an assistant professor of media arts and sciences. There are many misread problems if an RFID tag is occluded or misoriented, Ma notes. Also on the research team is Nicholas Selby, a mechanical engineering MIT graduate student.

The project aims to streamline storage-site tag reading with RFly, a drone-based RFID technology (Figure 1). "RFly builds a lightweight, airborne, full-duplex relay that amplifies and forwards an RFID reader's commands to an RFID tag, and also relays the RFID tag's response back to the reader," Ma says. "In doing so, RFly is able to extend today's battery-free network communication range by ten times and coverage area by 100 times."

Unlike conventional relays, RFly is designed to preserve the critical phase

Digital Object Identifier 10.1109/MSP.2018.2847567
Date of publication: 28 August 2018

and timing information of the signal at the physical layer, making it transparent to both the RFID reader and tag. “RFly can be readily integrated with today’s RFID infrastructure without the need for hardware and software modifications,” Ma states. Along with the range extension, RFly offers an accurate localization service. “It leverages the continuous drone flightpath to emulate a virtual antenna array, so the location of an RFID tag can be accurately pinpointed with an error smaller than 20 cm, even if an RFID tag is tens of meters away from the reader,” Ma explains.

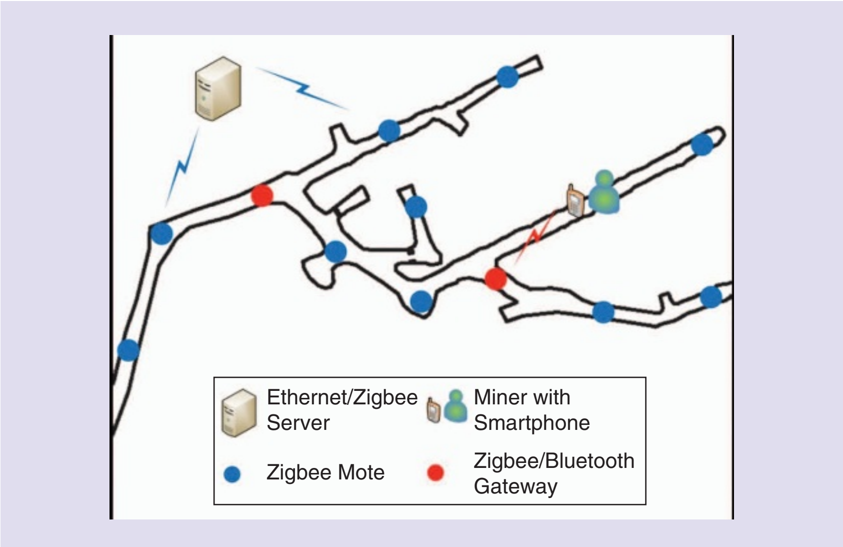
By extending the communication range by an order of magnitude, RFly promises to drastically reduce the density and deployment cost of today’s typical RFID reader infrastructure. According to Ma, flying continuously, a drone can gain multiple RFID tag perspectives to eliminate wireless blind spots, significantly reducing misreads caused by occlusion or misorientation. “The localization capability allows you to not only identify but locate misplaced items,” he adds.

Ma notes that signal processing is critical to the project’s success. “We leverage frequency shifting, window function, and bandpass filtering to cut the feedback loop of the self-leakage path in the relay circuit to avoid oscillation, and that allows greater signal amplification,” he explains. “We utilize matched filters to boost signal-to-noise ratio in the process of extracting signal amplitude and carrier phase,” he says. “We harness synthetic aperture radar equations in the process of calculating the RFID tag location using a mobile drone.”

Ma observes that the project’s algorithms go hand in hand with circuit design. “Our system presents a unique hardware–software cross-layer optimization, which allows us to achieve much better performance. Therefore, the algorithm compatibility to the hardware is one of the most important considerations.”

Because of the drone’s limited payload and battery capacity, weight and power consumption were also key design considerations. “We also need to handle unexpected interference and noise, so the robustness of the signal processing algorithm should also be taken into account,” Ma says. “Overall, it is very important to keep the choice of signal processing approaches concise and elegant.”

In experiments in the Media Lab involving tagged objects—many intentionally hidden to approximate the condition of merchandise heaped in piles on warehouse shelves—the system was able to localize tags with 19-cm accuracy while extending the range of the reader tenfold in all directions, or 100-fold cumulatively.



COLORADO STATE UNIVERSITY

FIGURE 2. A proposed system layout for underground mine monitoring, tracking, and communication.

Going underground

According to Colorado State University, the United States alone has 13,000 active mines. Keeping miners safe and productive is a major challenge given the fact that radio signals can’t penetrate deeply underground. A project led by university engineers aims to provide miners with a low-cost, high-fidelity communications system that bypasses global positioning systems, wireless, cellular, and other signals that are taken for granted above ground.

According to Prof. Sudeep Pasricha, chair of computer engineering at Colorado State University and principal investigator in the SmartMiner project, in the past decade, mine accidents have killed 40,000 mine workers worldwide, 500 of which were in the United States. “Unfortunately, the high cost of deploying a safety infrastructure encourages companies to meet only the minimum required safeguards,” Pasricha notes. “Mine safety demands a scalable, low-cost solution to enable sensing, communication, and tracking in underground mines to detect precursors to mishaps and also aid rescue efforts in the aftermath of an accident, such as a tunnel cave-in or an explosion.”

The project’s wireless cyberphysical framework incorporates standard smartphones and low-cost wireless sensing. It aims to eliminate the need for expensive handheld and communication equipment and give miners an extra edge of safety, whether for day-to-day communication or during an emergency. “The objective of our project is to devise, design, prototype, and test a fundamentally novel wireless cyber-physical framework of low-cost, energy-efficient, and reliable sensor nodes and commodity smartphones for monitoring, tracking, and communication,” Pasricha says.

The planned framework (Figure 2) offers a wireless network consisting of multiple low-cost stationary Zigbee or Bluetooth sensors deployed strategically throughout a mine, creating a mesh network that can connect with smartphones carried by miners. The precise placement of the fixed sensors is based on an analysis of how radio signals travel in an underground mine's complex, changing, and noisy environment. The researchers are also designing new software algorithms and filtering techniques developed for use on smartphones. When connected to the wireless mesh network, the researchers believe that the system will be able to accurately and efficiently calculate a miner's location within a mine, despite the highly unpredictable nature of wireless signals.

For underground localization, the system leverages data signals from several different sensors and wireless radios. Determining the heading angle—the angle a user is facing with respect to true North—is obtained by combining accelerometer, gyroscope, and magnetometer readings. The accelerometer provides the gravity vector, the magnetometer functions as a compass, and the gyroscope provides angular rotation speed. Angular rotation speed data signals are integrated over a time interval to determine the orientation of the mobile device, Pasricha says. Sensor data from all three sensors are then combined using Kalman filtering to obtain precise orientation that avoids both gyro drift and noisy orientation. "We can use this … data to determine the change in position from the current position," Pasricha states. "In a similar manner, we use various signal processing approaches to combine the fused inertial sensor data with wireless signal data to localize a user."

Signal processing is also used to assist with characterizing the environment. "For instance, when simulating an underground mine, we first require a digital representation of the tunnel system in the form of a triangular mesh," Pasricha says. Since mine tunnels are large and highly asymmetric, hand measurement and manual mesh building are out of the question. "Instead, we use lidar scanning to collect millions of measurements of the mine in the form of point cloud data," Pasricha explains. The data is used to construct a triangular mesh. "However, these data are noisy and only represent a discrete set of samples of a continuous object—the mine walls." The data also doesn't provide any details relating to the scanned surface's normal vectors or what that surface may look like between sample points. "We, therefore, draw on signal processing techniques, like principal component analysis, to extract properties like local curvature and surface normal from the point cloud data in a way that is very tolerant of noise," Pasricha says.

"The signal processing methods we use in our research evolved very organically with our work," Pasricha observes. "We have often worked iteratively, starting with a sample approach and only complicating it if we find that the simple technique is giving bad results." For instance, at one point, the researchers used uniform random sampling for ray initialization as part of their underground mine characterization studies because they felt it was fast and easy to implement. "But [we] found that this was far less effective than using Halton sampling, which, although slower to initialize, has improved [the] convergence rates of our simulator dramatically," Pasricha says. Other project participants include coprincipal investigator Prof. Branislav Notaros of Colorado State University and Prof. Qi Han of the Colorado School of Mines.

A new foundation

At Finland's Aalto University, researchers are addressing two key IoT challenges: device size and power consumption. Sayani Majumdar, an academy fellow in Aalto University's Department of Applied Physics, is leading a team targeting both issues, with the goal to make everyday IoT tags and sensors smaller and less power hungry.

The research is critical, Majumdar notes, because IoT will soon require the capability to process and store an unprecedented amount of data, resulting in a huge energy cost. "Moreover, current CMOS technology will soon be unable to undergo further miniaturization," she adds.

Majumdar's team is pegging its hopes on memristor technology, which they view as a viable and superior approach to replacing transistors. An abbreviation of the term *memory resistor*, a memristor is a two-terminal resistive switching device that can retain a memory of its last resistance state, even after being powered off.

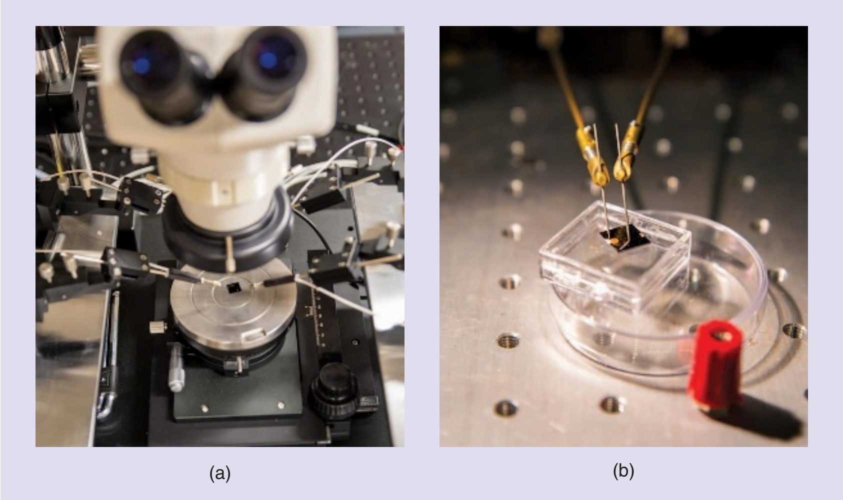
Memristors also have the ability to change their resistance states continuously rather than in binary zeros and ones. "This analog control of various current conduction states offers the opportunity of mimicking the human-brain-like activities, which is the goal of neuromorphic computing," Majumdar says. The capability promises to lead to a new generation of intelligent IoT devices.

The memristor technology developed by the researchers is a ferroelectric tunnel junction. The device, a nanometer-thick ferroelectric thin film squeezed between a pair of electrodes, can function with voltages of under 5 V. The technology is compatible with a wide range of electrode materials, including silicon, and can be manufactured in conventional production facilities. Tunnel junctions also offer the benefit of being able to retain data for over a decade without requiring a power source.

"Our devices work based on the amplitude, duration, and time delay between electrical impulses and also on the history of the previous state of the memristors," Majumdar says. "This is very similar to how the human brain processes information through electrical impulses, and, therefore, our devices are very suitable for neuromorphic computation." A key IoT need is ample data storage and processing at an affordable energy cost. "Memristor-based neuromorphic circuitry can provide a solution in this respect," Majumdar adds.

Once the junctions are in place and the researchers have ensured large and reproducible resistive switching in the devices, they apply different sets of input voltage pulses and record the output current to determine response activity (Figure 3). "The applied voltage magnitude range is from few millivolts to 5 V with a pulse shape either rectangular or arbitrary, comparable to that of biological systems with a pulsewidth of 20 ns–100 ms," Majumdar explains. "The output current can vary from microamperes to hundreds of picoamperes, based on the direction of ferroelectric polarization in our junctions."

Ferroelectric tunnel junctions are good memristors, Majumdar notes, exhibiting



TAPIO REINEKOSKI

FIGURE 3. The researchers' probe-station device, which is used to measure the electrical responses of the basic components for devices mimicking the human brain and advanced IoT applications. (a) The full instrument and (b) a closer view of the device connection where the tunnel junctions are on a thin film on the substrate plate.

very low current and a fast operational speed—on a nanosecond time scale—making them highly energy efficient. "Our solution is a greener alternative to the complex oxide ferroelectric-based devices proposed in recent times that require high temperature processing and contain hazardous metals like lead, barium, or bismuth," she explains.

The researchers are now working to take the technology to a higher level. "What we are striving for now is to integrate millions of our tunnel-junction memristors into a network on a 1-cm^2 area," Majumdar says. "We can expect to pack so many in such a small space because we have now achieved a record-high difference in the current between on and off states in the junctions, and that provides functional stability." The memristors could then begin performing various kinds of complex tasks, such as image and pattern recognition and even make decisions autonomously.

Author

John Edwards (jedwards@johnedwardsmedia.com) is a technology writer based in the Phoenix, Arizona area. Follow him on Twitter @TechJohnEdwards.

SP

Chenren Xu, Yan (Lindsay) Sun, Konstantinos (Kostas) N. Plataniotis,
and Nic Lane

Signal Processing and the Internet of Things

The notion of the Internet of Things (IoT) has emerged as a last-mile solution for connecting various cybertechnologies to our everyday life. It envisions a three-tier architecture where highly distributed and heterogeneous sensor data will be collected through a gateway and made available to the Internet to be readily accessible for a wide range of applications. Today, with ever-increasing types of IoT devices as well as the growing demand being placed on the end user, the sensing platform, and the computing and storage infrastructure, more is being asked of engineers, designers, and scientists.

Signal processing is playing a progressively substantial role in this domain, including such general topics as analyzing, summarizing, and protecting signals and information exchanged or shared by connected things. The diversity of these problems requires a more collaborative effort from engineers and scientists from a varied set of specialties; yet there is no single domain to publish and communicate this to the general community. The impact to society is massive, including such broad aspects as energy efficiency, security and privacy considerations, and big data applications. How will signal processing advance today's autonomous networked sensor/device into interconnected ones in an energy-efficient, secure, and privacy-preserving manner? How can we leverage today's pervasive cloud and network infrastructure to foster more intriguing applications with more demanding signal processing and machine-learning techniques? There are clearly new and emerging challenges that need to be addressed.

This special issue contains 13 articles, and our aim is to provide a comprehensive view of the main advances in the field through a number of tutorial-style articles as well as through contributions that emphasize the key topics of development in this area, both in terms of theory and applications.

IoT devices are expected to operate in ultralow-power or even battery-free situations so they are easily deployed and run autonomously for a long time. Given that communication is the energy bottleneck, backscatter communication, an emerging μW-level wireless communication paradigm, is gaining popularity as a suitable solution to fulfill such need. The article "Practical Backscatter Communication Systems for Battery-Free Internet of Things" by Xu, Yang, and Zhang, surveys the practical bistatic backscatter system with design considerations covering energy efficiency, bit rate, communication range, and deployment cost.

How can we leverage today's pervasive cloud and network infrastructure to foster more intriguing applications with more demanding signal processing and machine-learning techniques?

Next, in their article "The Art of Signal Processing in Backscatter Radio for μW (or Less) Internet of Things" Bletsas, Alevizos, and Vougioukas offer a review on other perspectives such as (non)coherent signal processing at the reader, both symbol and symbol by symbol as well as sequence based, with or without channel coding.

Retaining security and privacy on an IoT system becomes very challenging because of its restricted computation, memory, radio bandwidth, and battery resources for executing computationally intensive and latency-sensitive tasks. Xiao et al.'s article, "IoT Security Techniques Based on Machine Learning," presents a comprehensive review focusing on the machine-learning-based IoT authentication, access control, secure offloading, and malware detection schemes to protect data privacy.

In the article "Approaches to Secure Inference in the Internet of Things," Zhang, Blum, and Poor focus on signal processing approaches to the development of active cyberattacks in inferential sensor processing for the IoT using quantized data, while Chen, Kar, and Moura review the algorithms for a secure, distributed interference in their article "The Internet of Things." Zhou et al.'s article "Security and Privacy for the Industrial Internet of Things" presents a summary of efficient cryptography for industrial IoT endpoints, scalable key management, and system privacy issues.

Digital Object Identifier 10.1109/MSP.2018.2846838
Date of publication: 28 August 2018

The massive deployment of IoT devices brings new challenges and opportunities. The article by Liu et al., "Sparse Signal Processing for Grant-Free Massive Connectivity," outlines several key signal processing techniques that are applicable to the problem of massive IoT access, focusing primarily on advanced compressed sensing technique and its application for efficient detection of the active devices.

Our aim is to provide a comprehensive view of the main advances in the field through a number of tutorial-style articles.

For indoor deployment, Tushar et al.'s article, "Internet of Things for Green Building Management," shows that IoT devices can collectively extract high-level building occupancy information through simple and low-cost IoT sensors. The article also studies the impact of human activities on the energy usage of a building, which can be exploited to design energy-conservation measures to reduce the building's energy consumption. Nathan et al. review problems in the fields of smart home sensing, signal processing, analytics, and visualization that require solutions cognizant with the specific needs of the elderly in "A Survey on Smart Homes for Aging in Place."

In their article, "From Surveillance to Digital Twin," He, Guo, and Zheng survey and discuss the challenges and recent works toward data acquisition, human–machine-product interconnection, knowledge discovery and generation, and intelligent control, from sensing and networking to the analytics layer. In "Crowd-Based Learning of Spatial Fields for the Internet of Things," Arias-de-Reyna et al. survey the problem of estimating the spatial distribution of physical quantities (spatial fields) by taking an advantage of the pervasive diffusion of IoT mobile devices equipped with sensors collecting measurements related to the spatial field at different locations.

"Microlocation for Smart Buildings in the Era of the Internet of Things," by Spachos, Papapanagiotou, and Plataniotis, covers the challenges and examines some signal processing filtering techniques, such that microlocation-enabling technologies and services can be thoroughly integrated with an IoT-equipped smart building. Last but not least, network localization and navigation, a design framework for the development of scalable and distributed techniques for multisensor fusion in the IoT, is presented in "Efficient Multisensor Localization for the Internet of Things" by Win et al.

Acknowledgments

We thank the contributors for their comprehensive and interesting articles and Prof. Min Wu for her support in putting together this special issue. We would like to also extend our thanks to the reviewers for their detailed and insightful comments and to IEEE Signal Processing Society Publications Administrator Rebecca Wollman for her guidance along the way.

Meet the guest editors

Chenren Xu (chenren@pku.edu.cn) received his Ph.D. degree in electrical and computer engineering with highest honors from Rutgers University, New Brunswick, New Jersey. He has held a postdoctoral fellowship at Carnegie Mellon University, Pittsburgh, Pennsylvania, and visiting positions at AT&T Shannon Labs and Microsoft Research. He is an assistant professor in the School of Electrical Engineering and Computer Sciences and a member of the Center for Energy-Efficient Computing and Applications at Peking University, Beijing, where he has directed the software–hardware orchestrated architecture group since 2015. He is the recipient of a Samsung Best Article Award (2014), Best Paper Nominee of the Association for Computing Machinery (ACM) UbiComp (2014), and Best Poster Award of ACM SenSys (2011). His research interests include the intersection of wireless, system, and networking, with a current focus on high-mobility data networking, visible light backscatter communication, and wireless edge system.

Yan (Lindsay) Sun (yansun@ele.uri.edu) received her B.S. degree with highest honors from Peking University, Beijing, in 1998 and her Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park, in 2004. She joined the University of Rhode Island, Kingston, in 2004, where she is currently a professor in the Department of Electrical, Computer, and Biomedical Engineering. She was editor-in-chief of the IEEE Signal Processing Society's SigPort (2015–2017) and is an associate editor of *IEEE Transactions on Signal and Information Processing Over Networks*. She was an elected member of the Information Forensics and Security Technical Committee (2014–2016), served on the editorial board of *IEEE Security and Privacy Magazine* (2013–2015), and was as an associate editor of *IEEE Signal Processing Letters* (2013–2015). She was the recipient of National Science Foundation CAREER Award (2007) for her work on trust management, and she has received multiple other awards.

Konstantinos (Kostas) N. Plataniotis (kostas@ece.utoronto.ca) is a professor and the Bell Canada Chair in Multimedia with the Electrical and Computer Engineering Department at the University of Toronto, Canada. He is the founder and inaugural director of research for the Identity, Privacy, and Security Institute and has served as the director for the Knowledge Media Design Institute from January 2010 to July 2012 at the University of Toronto. He is a Registered Professional Engineer in Ontario and a fellow of the Engineering Institute of Canada. He has served as editor-in-chief of *IEEE*

Signal Processing Letters and as technical cochair of the IEEE 2013 International Conference on Acoustics, Speech, and Signal Processing (ICASSP). He was the IEEE Signal Processing Society vice president of membership from 2014 to 2016. He is the general cochair for 2017 IEEE GlobalSIP, the 2018 IEEE International Conference on Image Processing, and ICASSP 2021. He is a Fellow of the IEEE.

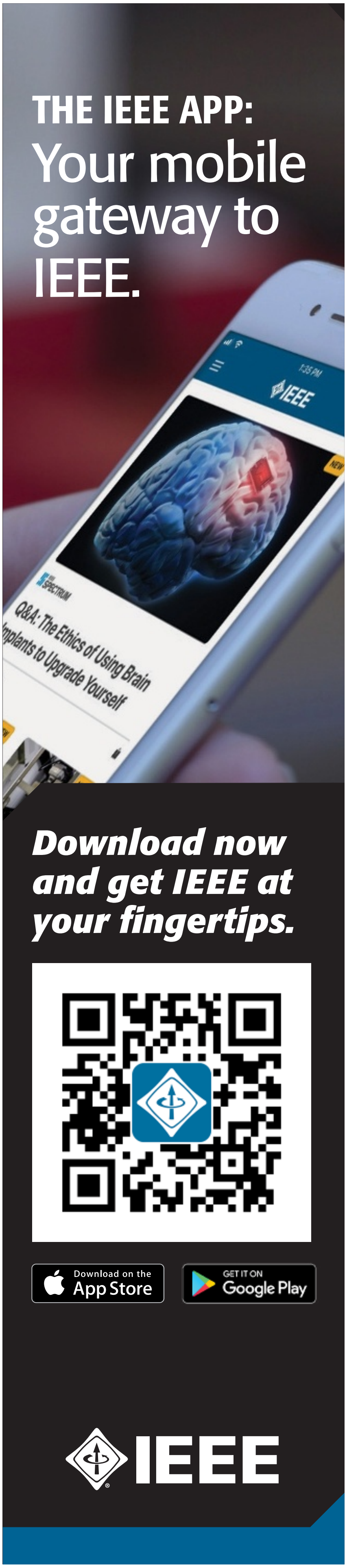
Nic Lane (niclane.lane@cs.ox.ac.uk) received his Ph.D. degree in 2011 from Dartmouth College, Hanover, New Hampshire. He is an associate professor in the Computer Science Department at the University of Oxford, United Kingdom. He is an experimentalist and likes to build prototype next-generation wearable and embedded-sensing devices based on well-founded computational models. His work has received multiple best paper awards, including one from the ACM/IEEE Conference on Information Processing in Sensor Networks 2017 and two from ACM UbiComp in 2012 and 2015, respectively.

FROM THE EDITOR

(continued from page 3)

end of the semester, and often coming close or conflicting with other conferences. The structure changed with each version, making each year an experiment. The conference did not always offer flexibility to the symposia organizers to have posters, oral sessions, and panels as they preferred. Technical committees were not heavily involved, as they are in ICASSP. There was no general session where people could submit papers that did not align with a symposium. I suspect it was a combination of reasons that led to GlobalSIP's demise. Could SPS have defined an internal team to work on launching the conference and develop the initial idea in a consistent way for the different editions of the conference?

I am left wondering about the connection between *SPM* and our conferences. Many feature articles are a by-product of a conference tutorial. This is not a coincidence, as we reach out to tutorial presenters to encourage submissions of feature articles. But what about special issues? Is there room for a hot topics symposium at ICASSP or the IEEE International Conference on Image Processing? Does it result from a natural clustering of papers at ICASSP? Or do we need another conference? I think GlobalSIP could have been the answer. After 2019, I do not know. There is so much value in special issues—I hope we can find a way to have such a feature in our conferences.


THE IEEE APP:
Your mobile gateway to IEEE.
IEEE
SPECTRUM
Q&A: The Ethics of Using Brain Implants to Upgrade Yourself
Download now and get IEEE at your fingertips.
Download on the App Store
GET IT ON Google Play
IEEE

Chenren Xu, Lei Yang, and Pengyu Zhang

Practical Backscatter Communication Systems for Battery-Free Internet of Things

A tutorial and survey of recent research

INTERNET OF THINGS—ISTOCKPHOTO.COM/IAREMENKO
CIRCUITS—IMAGE LICENSED BY INGRAM PUBLISHING

Backscatter presents an emerging ultralow-power wireless communication paradigm. The ability to offer submilliwatt power consumption makes it a competitive core technology for Internet of Things (IoT) applications. In this article, we provide a tutorial of backscatter communication from the signal processing perspective as well as a survey of the recent research activities in this domain, primarily focusing on bistatic backscatter systems. We also discuss the unique real-world applications empowered by backscatter communication and identify open questions in this domain. We believe this article will shed light on the low-power wireless connectivity design toward building and deploying IoT services in the wild.

Overview of backscatter communication

The vision of the IoT promises a world where sensors and actuators are ubiquitous and interconnected so that we can better understand and control the surrounding world. One critical challenge toward this vision is to build such devices that can be easily deployed and run autonomously for a lengthy duration. Backscatter communication, an emerging microwatt-level wireless communication paradigm, is gaining popularity as a suitable solution to fulfill such a need.

The principle of backscatter communication is similar to that of the heliograph shown in Figure 1. People have been using mirrors to reflect sunlight for communication for a long time, and this method is especially important when there is no source of energy like a campfire or a flashlight. By flipping the mirror, the sender can signal the remote target by controlling the presence of reflected light using Morse code. For backscatter communication, the same reflecting while manipulating process is applied on radio-frequency (RF) signals. At a high level, the system model of backscatter communication is shown in Figure 2. A special device called a *backscatter tag* reflects the incoming excitation signal emitted by a nearby (carrier) transmitter. At the same time, it selectively changes the amplitude, frequency, and/or phase of the signal for modulation. The backscattered signal is then captured by a receiver and piped through a signal processing engine to extract information

Digital Object Identifier 10.1109/MSP.2018.2848361
Date of publication: 28 August 2018

injected by the backscatter tag. Note that the transmitter and the receiver were previously integrated in the conventional or monostatic backscatter system [e.g., RF identification (RFID) reader [1]] but are separated in bistatic backscatter system designs. Specifically, the transmitters can be available ambient RF sources [e.g., TV or frequency modulation (FM) radio towers, cellular base stations, and Wi-Fi access points (APs)] from anywhere. This new modular design introduces the following intrinsic properties for performance enhancement.

1) *Temporal flexibility*: In many sensing applications, it is important for the backscatter tag (as a sensor node) to transmit as soon as the sensory data are available. When excitation signals can potentially come from multiple sources, the tag has more time slots for data transmission instead of waiting for protocol-constrained interrogation from a single reader.
2) *Spatial flexibility*: The coverage of excitation signals is vital to the performance of backscatter communication. Being decoupled from the receiver, the transmitter(s) can be strategically placed in optimal locations to balance the scalability and performance for backscatter tags.
3) *Technology flexibility*: Bistatic backscatter system design presents a general and technology-independent communication paradigm that allows a variety of excitation signals and modulation schemes to be used in situ. Ambient RF sources, e.g., Wi-Fi signals, can be used to make the backscatter technology immediately deployable, because there are commodity Wi-Fi transmitters (e.g., APs) and receivers everywhere.

The main advantage of backscatter communication is energy efficiency. Compared with conventional wireless technologies, such as Wi-Fi (tens of milliwatts), Bluetooth/Bluetooth Low Energy (several milliwatts), and long-term evolution (LTE) (hundreds of milliwatts), the power consumption of backscatter communication is more than 1,000 times smaller. The key to realize such power reduction is that the procedure of radio signal generation, i.e., the most power-consuming block in radio communicators, is offloaded to the powered transmitter and thus is not present in a backscatter tag. In addition, signal amplification and processing are also delegated to the transmitter. This creates an asymmetric design consisting of a fat transmitter/receiver and a thin backscatter tag. The ultralow-power nature of such a design makes it feasible for backscatter tags to be battery-free by utilizing today's energy-harvesting techniques, such as solar/light, mechanical motion/vibration, thermoelectric effect, and electromagnetic radiation [2], [3] with the limited form factor (square centimeter) of common IoT devices. Apart from the benefit of energy efficiency, this asymmetric design also means simpler hardware design, smaller form factor, and lower cost of the tag. The idea of such an asymmetric design can also be extended to provide a more energy-efficient wireless link for day-to-day use of mobile devices [4]. This design is making backscatter a competitive solution for IoT devices, and it is an important step toward realizing large-scale IoT application deployment in the wild.

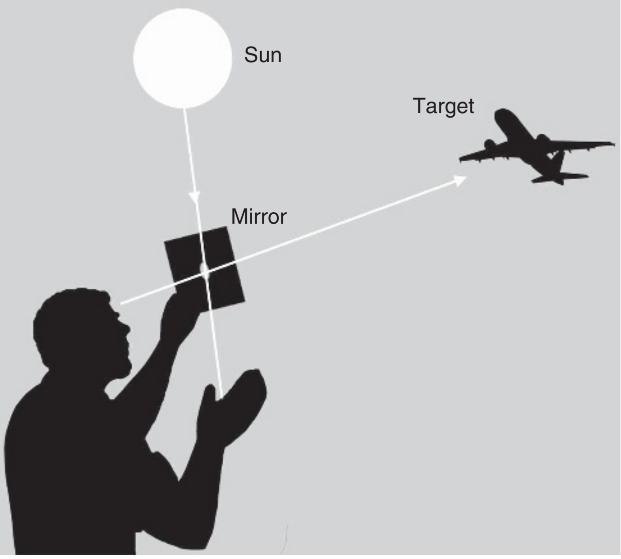


FIGURE 1. A heliograph, a simple but effective instrument that signals by flashes of sunlight reflected by a mirror for instantaneous optical communication over long distances. The flashes are modulated by momentarily pivoting the mirror or by interrupting the beam with a shutter.

Tutorial on backscatter communication

Backscatter basics

While various backscatter communication technologies are available, all of them are based on the same or similar model and techniques, which is to enable backscatter tags to reflect an incoming RF signal and at the same time modify and modulate the signal for secondary transmission, or backscatter. The core idea of modifying and reflecting the RF signal is impedance mismatching. On a backscatter tag, such discontinuity can be implemented by connecting an antenna of impedance $Z_A = |Z_A|e^{j\theta_A}$ to a load of impedance $Z_L = |Z_L|e^{j\theta_L}$. The

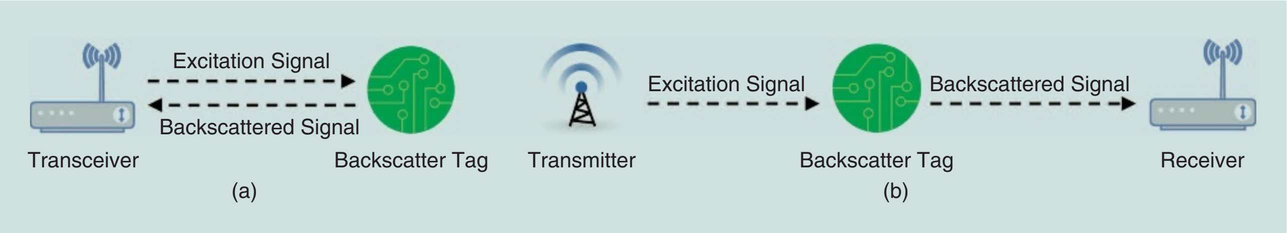


FIGURE 2. The (carrier) transmitter is separated from the receiver in the (modern) bistatic backscatter system design, in comparison to the (conventional) monostatic backscatter system model: (a) the monostatic backscatter system and (b) the bistatic backscatter system.

reflection coefficient of the backscatter tag circuit Γ_T can be calculated as (1), where $|\Gamma_T|$ and θ_T are given in (2) and (3):

$$\Gamma_T = \frac{Z_L - Z_A}{Z_L + Z_A} = 1 - \frac{2|Z_A|}{|Z_A| + |Z_L| e^{-j(\theta_A - \theta_L)}} = |\Gamma_T| e^{j\theta_T}, \quad (1)$$

$$|\Gamma_T| = \frac{|Z_A|^2 + |Z_L|^2 - 2|Z_A||Z_L|\cos(\theta_A - \theta_L)}{|Z_A|^2 + |Z_L|^2 + 2|Z_A||Z_L|\cos(\theta_A - \theta_L)}, \quad (2)$$

$$\theta_T = \arctan\left(\frac{2|Z_A||Z_L|\sin(\theta_A - \theta_L)}{|Z_A|^2 - |Z_L|^2}\right). \quad (3)$$

As the name suggests, the reflection coefficient Γ_T describes the ratio of the complex amplitudes of the incoming signal $S_{\text{in}}(t)$ and the reflected signal $S_{\text{out}}(t)$. For simplicity, we define the incoming signal $S_{\text{in}}(t)$ as a sine wave, as shown in (4):

$$S_{\text{in}}(t) = A_{\text{in}} e^{j(2\pi f_{\text{in}} t + \theta_{\text{in}})}. \quad (4)$$

Note that we do not lose generality here by replacing arbitrary $S_{\text{in}}(t)$ with a sine wave because any $S_{\text{in}}(t)$ can be regarded as a collection of sine waves using Fourier transform. Given the definition of $S_{\text{in}}(t)$ and the reflection coefficient Γ_T, the reflected signal $S_{\text{out}}(t)$ is calculated and shown in (5):

$$S_{\text{out}}(t) = \Gamma_T \cdot S_{\text{in}}(t) = |\Gamma_T| A_{\text{in}} e^{j(2\pi f_{\text{in}} t + \theta_{\text{in}} + \theta_T)}. \quad (5)$$

We can see that the backscatter tag is able to control $S_{\text{out}}(t)$ by controlling Γ_T, which is derived from the impedance of the antenna Z_A and the load Z_L. As a result, we can adjust Z_L to change the value of Γ_T for modulating $S_{\text{out}}(t)$. Unlike conventional radio communication systems that can directly change the amplitude, frequency, and phase of $S_{\text{out}}(t)$ for modulation, in backscatter systems, $S_{\text{out}}(t)$ can only be manipulated by changing Γ_T. As shown in (5), Γ_T effectively applies a phase shift of θ_T and an attenuation of $|\Gamma_T|$ to the incoming signal $S_{\text{in}}(t)$, where

$$A_{\text{out}} = |\Gamma_T| \cdot A_{\text{in}} \quad \text{and} \quad \theta_{\text{out}} = \theta_{\text{in}} + \theta_T. \quad (6)$$

By selecting between different Γ_T values, the backscatter is able to toggle the reflected signal $S_{\text{out}}(t)$ among a set of amplitudes and phases. To implement it on a backscatter tag, a commodity electronic component called an *RF switch* [5] is used. An RF switch is able to route the high-frequency signal through different transmission paths. On the backscatter tag, the RF switch is used to connect the antenna to RF loads with different impedance and switch between them. A low-power microcontroller unit (MCU) or a field-programmable gate array (FPGA) is then used as a controller to control the RF switch. As a result, the backscatter tag is able to adjust Γ_T and thus control A_{out} and θ_{out}. On top of these basic operations, we develop the backscatter tag design taxonomy based on the following two aspects: frequency shifting (FS) or not, and digital or analog modulation. The comparison of different backscatter systems surveyed in this article based on this taxonomy is summarized in Table 1.

FS

One of the major differences between backscatter tag designs is whether the tag has the ability to change the frequency of the backscattered signal.

Backscatter without FS

Because the backscatter tag is able to adjust A_{out} and θ_{out} by changing Γ_T, it can use a set of different amplitudes, phases, or their combinations to represent data. For digital data, $\Gamma_T(t)$ follows the function shown in (7):

$$\Gamma_T(t) = \begin{cases} \Gamma_0 & \text{when transmitting symbol 0} \\ \Gamma_1 & \text{when transmitting symbol 1} \\ \cdots & \\ \Gamma_n & \text{when transmitting symbol } n \end{cases}, \quad (7)$$

where $\Gamma_0, \Gamma_1, \ldots, \Gamma_n$ are discrete values producing different $S_{\text{out}}(t)$. For analog data, special circuits are designed to convert the value of the input data, such as voltage, to Γ_T, so that A_{out} and θ_{out} change accordingly. By doing so, the backscatter tag is able to perform amplitude-shift keying, amplitude modulation (AM), phase-shift keying (PSK), phase modulation (PM), or their combinations.

Non-FS backscatter tags use a simple design that maps the input data directly to the amplitude and phase of $S_{\text{out}}(t)$.

Table 1. A comparison of different backscatter tag designs.

Modulation	FS	Examples	Figure	Advantages	Disadvantages
Digital	No	BackFi [6] Ambient backscatter [7] Wi-Fi backscatter [8]	Figure 3	Simple design	May cause self-interference
	Yes	Passive Wi-Fi [9] HitchHike [10] LoRea [11] Interscatter [12]	Figure 4	More flexible in controlling the backscattered signal and supporting more modulation schemes	May generate unwanted sidebands and cause interference when $\Gamma_T(t)$ is not pure sinusoidal
Analog	No	Battery-free cell phone [13] Hybrid backscatter [5]	Figure 5	Energy efficient for analog data	Not applicable to all kinds of input data and may suffer from self-interference
	Yes	LoRa backscatter [14] FM backscatter [15]	Figure 6	Energy efficient for analog data, supports FM and CSS modulation	Not applicable to all kinds of input data and may cause interference when $\Gamma_T(t)$ is not purely sinusoidal

However, this design can cause problems on receiving $S_{\text{out}}(t)$ because $S_{\text{in}}(t)$ (emitted by the carrier transmitter) and $S_{\text{out}}(t)$ (emitted by the backscatter tag) will potentially interfere with each other at the receiver. In addition, such a design limits the usage of frequency-related modulation schemes, such as FM and FS keying (FSK).

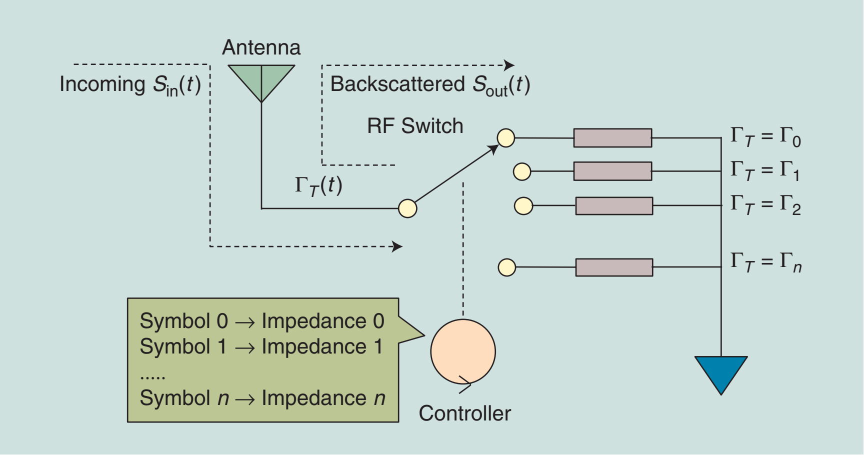


FIGURE 3. A typical backscatter tag that uses digital modulation without FS.

Backscatter with FS

To change the frequency of the reflected signal $S_{\text{out}}(t)$, the backscatter tags change Γ_T over time so that $\Gamma_T(t)$ is or approximates a sine wave $|\Gamma_T|e^{j(2\pi f_T t+\phi_T)}$. In this case, $S_{\text{out}}(T)$ can be calculated as (8):

$$S_{\text{out}}(t) = \Gamma_T(t)\cdot S_{\text{in}}(t) = |\Gamma_T|\,A_{\text{in}}e^{j(2\pi(f_{\text{in}}+f_T)t+(\theta_{\text{in}}+\phi_T))}. \tag{8}$$

As a result, $S_{\text{out}}(t)$ is frequency shifted by f_T, phase shifted by ϕ_T from $S_{\text{in}}(t)$, and attenuated by $|\Gamma_T|$.

While the sinusoidal $\Gamma_T(t)$ can actually be a sine wave signal, many backscatter tag systems use digital signals like a square wave to approximate a sine wave. For simplicity, here we define $\Gamma_T(t)$ as a square wave, shown in (9):

$$\Gamma_T(t) = \frac{A_T}{2}\text{sgn}(\sin(2\pi f_T t+\phi_T)) + \frac{A_T}{2} = \sum_{k=0}^{\infty}\gamma_k(t). \tag{9}$$

Note that the same method that we use to analyze the square wave can be applied to other types of digital signals as well. In this case, Γ_T is toggled back and forth between 0 and A_T, with frequency f_T and phase ϕ_T. $\Gamma_T(t)$ can be expanded into a series of $\gamma_k(t)$ elements using Fourier transform. The definition of $\gamma_k(t)$ is provided in (10):

$$\gamma_k(t) = \begin{cases} \frac{A_T}{2} & k=0 \\ \frac{A_T}{k\pi}\left(e^{j\left(-2k\pi f_T t-k\phi_T+\frac{\pi}{2}\right)} + e^{j\left(2k\pi f_T t+k\phi_T-\frac{\pi}{2}\right)}\right) & k=1,3,5,\dots\,. \\ 0 & k=2,4,6,\dots, \end{cases} \tag{10}$$

when k is a positive odd number and γ_k is a pair of sine waves, which are desired by FS backscatter tags. In this case, γ_k is able to create a pair of sidebands in $S_{\text{out}}(t)$, as shown in (11):

$$\begin{aligned}&\gamma_k(t)S_{\text{in}}(t)\\ &= \frac{A_{\text{in}}A_T}{k\pi}\left(e^{j\left(2\pi(f_{\text{in}}-kf_T)t+\theta_{\text{in}}-k\phi_T+\frac{\pi}{2}\right)} + e^{j\left(2\pi(f_{\text{in}}+kf_T)t+\theta_{\text{in}}+k\phi_T-\frac{\pi}{2}\right)}\right).\end{aligned} \tag{11}$$

The sidebands are frequency shifted by $\pm kf_T$, phase shifted by $\pm(-k\phi_T+\pi/2)$ from S_{in}, and attenuated by $(A_T/k\pi)$. Note that, as long as $\Gamma_T(t)$ is not pure sinusoidal, the tag will produce multiple sidebands in $S_{\text{out}}(t)$, of which only one is used for transmitting data, i.e., other unsed sidebands may cause interference to surrounding wireless devices. Although there have been proposals on removing some of those sidebands [14], a solution to completely eliminate the interference in a low-power manner is yet to be developed.

To implement an FS backscatter, the tag needs to generate $\Gamma_T(t)$. As mentioned previously, many tags generate a square wave to approximate a sinusoidal. In this case, an RF switch is used to connect different RF loads to the antenna, and an FPGA or MCU toggles the RF switch between Γ_T and 0 at frequency f_T and phase ϕ_T to generate the square wave. Other types of $\Gamma_T(t)$ can also be generated in a similar way. In real-world implementations, we observe that FPGA is often preferred over an MCU as the controller because it consumes less energy when running at the same clock rate. A Freescale Kinetis low-power MCU running at 50 MHz, e.g., can use up to 23 mW, while the Microsemi IGLOO FPGA in HitchHike [10] at the same clock rate consumes fewer than 2 mW.

Digital/analog modulation

Similar to conventional communication systems, the modulation process of backscatter signals can also be digital or analog.

Digital modulation

When a backscatter tag uses digital modulation, it maps symbols to different $S_{\text{out}}(t)$ waveforms that vary in frequency, amplitude, or phase. To do so, the backscatter tag generates $\Gamma_T(t)$ that changes with the symbol to be transmitted by having a controller (e.g., an MCU or FPGA) to switch between the finite set of discrete states (see Figures 3 and 4).

Analog modulation

In analog modulation, $S_{\text{out}}(t)$ changes continuously. It is achieved by converting the input data to the frequency, amplitude, and phase of $\Gamma_T(t)$ using dedicated analog circuits. The input voltage can control the output frequency of a voltage-controlled oscillator (VCO), e.g., and, hence, the frequency of a sine wave $\Gamma_T(t)$ signal. In this case, the backscatter tag creates an $S_{\text{out}}(t)$ that is frequency modulated. Analog backscatters vary in the method to control $\Gamma_T(t)$, which depends on the source and the properties of input data. By using analog modulation, the backscatter tag is able to directly convert the analog

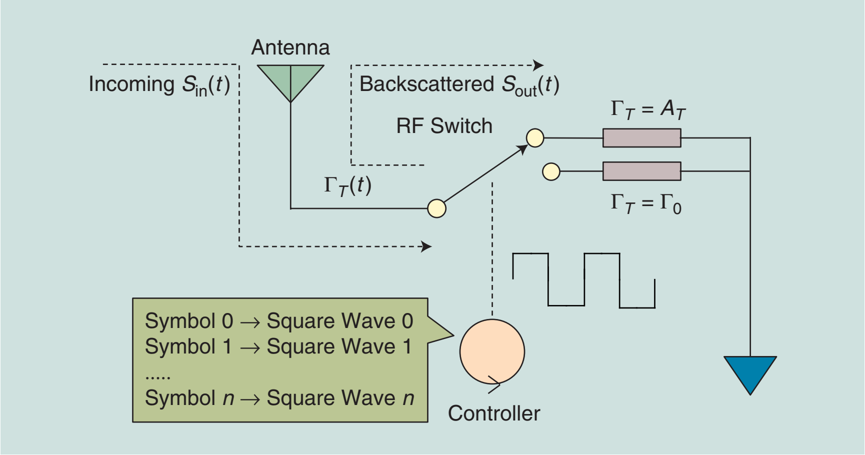


FIGURE 4. A typical backscatter tag that uses digital modulation with FS.

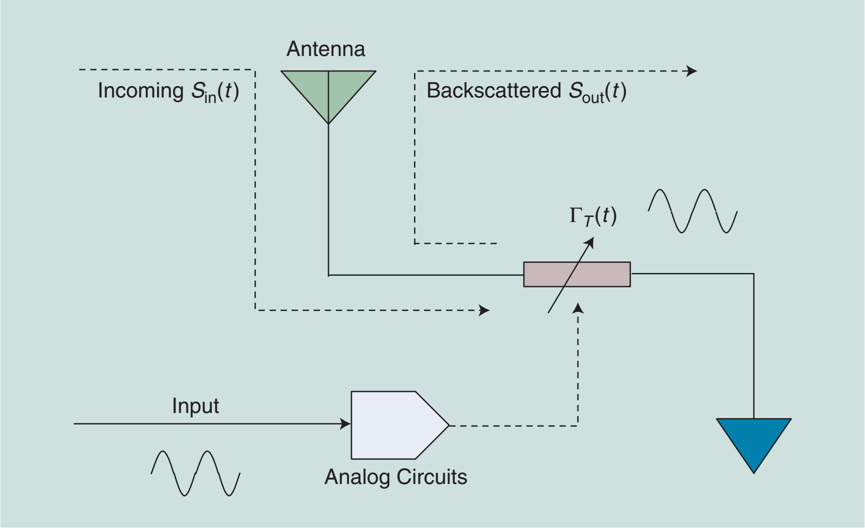


FIGURE 5. A typical backscatter tag that uses analog modulation without FS.

input to analog-modulated $S_{out}(t)$ without processing it in the digital domain, which eliminates the need for computational components, such as MCUs or FPGAs, as well as analog-to-digital converters (ADCs) (for purposes of sensing), which are power hungry especially when processing high-bandwidth data streams, such as audio and video. However, analog modulation is not able to deal with all kinds of input data, especially when the data are already in digital format. A case-by-case analog circuit design has to be developed to correlate the data source to $\Gamma_T(t)$ so that the input data modulate $S_{out}(t)$ (see Figure 5).

State-of-the-art backscatter systems

Despite the simplicity of backscatter operation in principle, there is a long way to go toward delivering a practical (bistatic) backscatter communication system good enough for real-world IoT applications. The research community has identified four main challenges of backscatter communication: energy efficiency, bit rate, communication range, and deployment cost. Motivated by the promising advantages of backscatter communication, research activities and efforts to tackle those problems have flourished in recent years. We summarize the key performance of the surveyed research projects in Table 2 and elaborate on them based on these performance metrics in the rest of this section.

Energy efficiency

The ultralow-power nature of backscatter communication makes it promising to run IoT applications in a battery-free manner. In extreme cases, applications would even be in need of always-on communication, which makes RF signals (rather than solar, vibration, and so forth) as the ambient energy source probably the only choice for energy harvesting. Consequently, there is great incentive to build backscatter communication systems that can be powered entirely by harvesting ambient RF energy. However, typical RF harvesting efficiency is as low as 18.2% when ambient RF signal strength is –20 dBm and only 0.4% at –40 dBm [16]. Such a low-energy budget poses great challenges to the hardware and system design.

Recent years have seen projects exploring the design and implementation toward this goal. Ambient backscatter [7] presents the first backscatter communication system that runs solely on energy harvested from ambient RF signals, such as TV and cellular towers. In this system, no customized carrier transmitter is deployed. The highlight of this system is that the backscatter tags have two-way communication capabilities and can talk to each other directly. Data injection on the backscatter tag is realized by controlling the RF switch to reflect or absorb the ambient RF signal. It reflects signals when transmitting 1, e.g., and absorbs signals when transmitting 0. Such operation creates a difference in the RF energy detected by a nearby backscatter tag. However, it is challenging to detect such a change on the receiving backscatter tag, because the ambient RF signal is already weak and has been modulated to convey other information, such as TV. To reliably decode data on backscatter tags, a two-step decoding circuit is designed. First, the received RF signal passes through an envelope detector and an averaging circuit to get the current RF energy level. Then it passes through a threshold comparator to decide whether or not the transmitting tag is a reflecting RF signal. The threshold here is computed by taking a long-term average of the signal. This system achieves a throughput of 1 kilobit/s over 0.76 and 0.46 m in outdoor and indoor environments, respectively.

Table 2. A performance comparison of state-of-the-art backscatter communication systems.

		Maximum Bit Rate		Range		Deployment	
Name	Minimum Power	Bit Rate	Distance	Transmitter to Tag	Tag to Receiver	Transmitter	Receiver
BackFi [6]	N/A	5 megabit/s	1 m	7 m	7 m	Ambient Wi-Fi	Software-defined radio
Ambient backscatter [7]	0.79 μW	10 kilobit/s	0.4 m	N/A	2.5 m	Ambient TV	Customized hardware
Wi-Fi backscatter [8]	9.65 μW	1 kilobit/s	N/A	N/A	2.1 m	Commodity Wi-Fi	Commodity Wi-Fi
Passive Wi-Fi [9]	14.5 μW (IC)	11 megabit/s	N/A	3.7 m	16.8 m	Customized hardware	Commodity Wi-Fi
HitchHike [10]	33 μW (IC)	300 kilobit/s	34 m	1 m	54 m	Commodity Wi-Fi	Commodity Wi-Fi
LoRea [11]	70 μW	197 kilobit/s	175 m	1 m	3.4 km	Patched commodity	Customized hardware
Interscatter [12]	28 μW (IC)	11 megabit/s	N/A	0.9 m	27.4 m	Commodity BLE	Commodity Wi-Fi/Zigbee
Battery-free cell phone [13]	3.48 μW	N/A	N/A	15.2 m	15.2 m	Customized hardware	Customized hardware
LoRa backscatter [14]	9.25 μW (IC)	37.5 kilobit/s	N/A	5 m	2.8 km	Customized hardware	Commodity LoRa
FM backscatter [15]	11.07 μW (IC)	3.2 kilobit/s	4.9 m	N/A	18.3 m	Ambient FM	Commodity FM

Note: In the "Range" column, the operating range of a bistatic backscatter system consists of two parts: the distance between transmitter and tag and the distance between tag and receiver. Here, we provide the maximum tag-to-receiver distance and the corresponding transmitter-to-tag distance. Under "Maximum Bit Rate," the "Distance" is the tag-to-receiver distance when achieving the maximum bit rate. For passive Wi-Fi and all numbers marked "IC", the number in the "Minimum Power" column is the simulation result of an IC design. IC: integrated circuit; BLE: Bluetooth/BLE.

The battery-free cell phone project [13] presents another battery-free backscatter system that is able to sense and transmit voice as well as receive and actuate audio. In this system, a battery-free backscatter tag works with a nearby customized base station (connected to cellular networks) to send and receive audio data. The key contribution of this system is the design and use of analog backscatter, where analog data, such as a wave signal, are directly backscattered without being converted and processed in the digital domain. This new design is even more energy efficient than the (conventional) digital backscatter design, because all of the digital computational components, such as the FPGA and ADC/digital-to-analog converter (DAC) (for converting sound to and from a digital signal), that potentially become the bottleneck of battery-free operation are eliminated to save energy. To convert the input audio to the impedance of the antenna, a special component called an *electret microphone* is used. Inside the electret microphone, there is a junction-gate field-effect transistor (JFET). In this design, the JFET is configured to work in its triode region and acts as a voltage-to-impedance converter. So the backscatter tag modifies the incoming signal according to the change of the audio voltage and eliminates the need for digital circuits. To receive and actuate downlink audio without the need for digital components, AM audio is transmitted from the base station to the backscatter tag, passes through an AM demodulator, and is directly fed into an earphone. To allow the exchange of control commands between the backscatter tag and the base station, the backscatter tag also has the ability to transmit and receive digital packets, but this functionality is only used when initiating and ending a call and thus does not use much energy overall. The device can work on RF energy harvesting when the cellular base station is 9.4 m away.

For other systems, the energy efficiency is traded for other goals. Passive Wi-Fi [13] and BackFi [6], e.g., implement differential quadrature PSK (DQPSK) and 16-PSK, respectively, to support 10+ megabits/s bit rate. Such high-speed baseband processing consumes much energy when compared to the cases where the RF switch toggles at a much lower speed [7] and the baseband processing is done in analog [13]. LoRa backscatter [14] trades power consumption for communication range, where much energy is used to generate chirp spread spectrum (CSS) modulation. HitchHike [10] and FreeRider [17] require precise synchronization with Wi-Fi packets in the air to provide compatibility with existing wireless protocols. Such synchronization needs low-delay RF energy detectors, which consume more power than the passive RF energy detectors used in other systems.

Bit rate

There are many IoT applications that require a wireless link faster than several kilobits per second. A smart speaker that continuously streams user voice to the cloud, e.g., needs a bit rate of more than 70 kilobit/s. Vision-based devices like security cameras can easily take up more than 1 megabits/s when in active operation. The bit rate requirement is even higher when considering emerging applications like augmented and virtual reality. Conventional radio technology tackles this problem by adopting advanced modulation schemes to make more efficient use of the available channel bandwidth, such as channel bonding and carrier aggregation. However, it is difficult for backscatter communication to use similar solutions. First, it is nontrivial to implement very complex modulations on a backscatter tag due to its limited capability in manipulating the RF signal and performing baseband processing. New techniques must be developed to enable more efficient use of the spectrum. Second, high throughput usually requires high-performance electronic components and fast computation, which further increases the energy consumption. As a result, it is nontrivial to support high-throughput applications atop backscatter communication systems, and a careful tradeoff must be made to balance throughput and per-bit energy consumption.

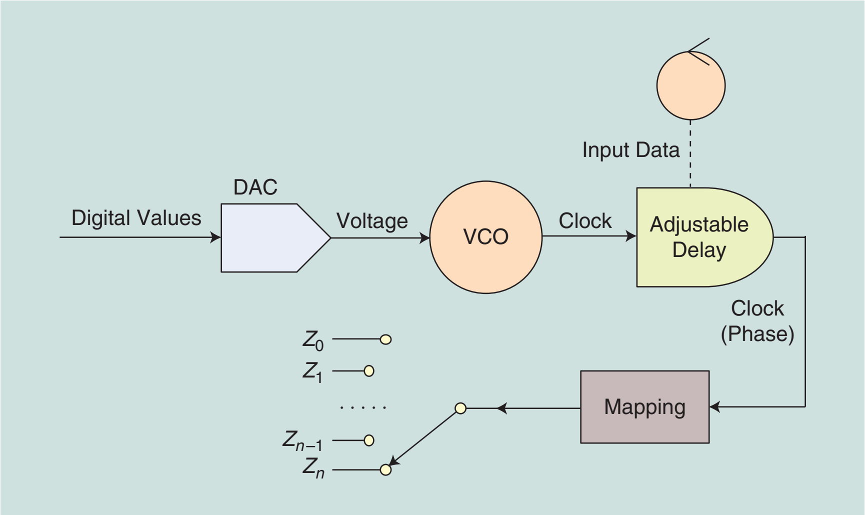


FIGURE 6. The design of the LoRa backscatter tag. The DAC and VCO are used to perform CSS modulation in a low-power manner. By replacing the SPDT RF switch in regular backscatter tags with a single-pole multithrow switch, a LoRa backscatter tag can cancel most of the sideband interference.

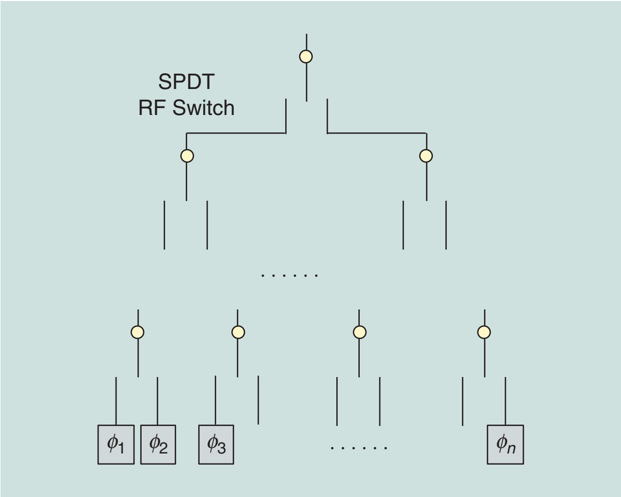


FIGURE 7. The design of a PM circuit in BackFi. Multiple SPDT switches are connected into a binary tree, and each leaf of the tree is connected to an RF delay line to provide precise phase shift of the incoming signal.

BackFi [6] presents a high-throughput Wi-Fi backscatter design that achieves 1,000 times higher throughput than the previous Wi-Fi backscatter system [8]. In this system, a customized Wi-Fi AP that is transmitting normal Wi-Fi packets to surrounding clients acts as both the backscatter transmitter and the receiver. The highlight of this system is that a novel backscatter tag is designed to provide efficient PMs, such as 16-PSK. Self-interference cancellation technology is used to allow the customized Wi-Fi AP to receive a reflected signal from backscatter tags while transmitting to surrounding Wi-Fi clients. In addition, an advanced decoder is designed to decode backscatter on wide-band signals, such as Wi-Fi. PM on backscatter tag is done by using a set of single-pole double-throw (SPDT) RF switches connected together to form a binary tree, as shown in Figure 7. RF delay lines of different length are connected at the tree leaves, providing different phase shifting, which is necessary to form the PSK constellation. Self-interference cancellation is done with the technology used in full-duplex radios. When a backscatter tag detects an incoming Wi-Fi signal, it waits for a short period of time before modulating the Wi-Fi signal, so that the self-interference cancellation circuits can use this time to estimate the channel between the backscatter tag and the receiver to avoid canceling actual backscattered data. This system is able to achieve 5 megabits/s at a range of 1 m.

Another backscatter system that provides high throughput is passive Wi-Fi [9], where the backscatter tag generates valid Wi-Fi 802.11b transmission at all bit rates, including the highest 11 megabits/s. In this system, a customized transmitter is used to transmit a single tone at a frequency out of Wi-Fi channels as the excitation signal, and a backscatter tag modulates this signal into valid Wi-Fi packets. All Wi-Fi-enabled devices can receive the transmission from the backscatter tag. One key contribution of this system is the design and implementation of a backscatter tag that performs baseband processing and modulates the incoming single-tone signal into 802.11b packets in a low-power manner. First, the backscatter tag generates a baseband of Wi-Fi packets following the specification of 802.11b. It then uses the techniques explained in the “Tutorial on Backscatter Communication” section to transform the single-tone signal. Specifically, it uses phase modification to perform differential binary PSK (DBPSK) and DQPSK, which are the two modulations used in 802.11b. It then uses frequency modification to move the signal to the center of the Wi-Fi channel. Another contribution of this system is a full network stack to enable multiple backscatter tags to share the channel and provide acknowledgment and rate adaption. Such a network stack requires a downlink from the transmitter to the backscatter tag. The downlink is encoded using on/off keying (OOK) and is decoded using a low-power RF energy detector on the backscatter tag.

Bit rate can be traded for other purposes as well. HitchHike [10] and FreeRider [17] backscatter valid Wi-Fi signals into valid Wi-Fi signals. To do so, they need to keep individual Wi-Fi symbols unbroken after being backscattered. Therefore, the smallest unit of modification is the size of a Wi-Fi symbol, which is 1 μs for HitchHike and even longer for FreeRider because of orthogonal frequency-division multiplexing (OFDM), which fundamentally limits the throughput. Ambient backscatter [7] trades throughput for power consumption. It uses the simple modulation scheme where the backscatter tag reflects the RF signal when transmitting 1 s and absorbs signal when transmitting 0 s. The bit rate is limited to enable successful decoding on the receiver. LoRa backscatter [14] trades throughput for communication distance. It uses the modulation schemes of LoRa, which is a low-rate long-range communication technology in nature.

Communication range

Outdoor IoT services often require a long communication range up to several kilometers. Although this requirement is already challenging to almost all wireless technologies, it is particularly hard to achieve in backscatter systems, primarily because of the higher path loss compared to conventional wireless communication. To be more specific, there can be significant loss when the signal is being reflected at the backscatter tag. Our experiments found that this attenuation could be up to 30 dB. In addition, radio signals experience path loss twice (from the transmitter to tag and from tag to receiver) instead of once. For conventional wireless technologies, this problem can be dealt with by either increasing the transmission power or adopting special modulation schemes, such as CSS used in LoRa. However, for backscatter, both solutions are not readily applicable. First, the transmission power may have to be increased to compensate for the extra attenuation in backscatter systems. Note that the transmission power has to be at least quadrupled when the total distance doubles, which can easily break through the limitations imposed by the electronic component, circuit design, power budget, and government regulations. Second, a backscatter tag can only perform limited processing on the incoming signal due to the power budget. As a result, special modulation schemes can be hard to implement on ordinary backscatter tags, which calls for special techniques to be developed to overcome this problem. All those issues make it challenging to build a long-range backscatter communication system.

When considering the problem of communication range, it should be noted that there are two distances involved in a backscatter communication system. The first is the distance from the transmitter to the backscatter tag (transmitter distance), which determines the range of the area where the backscatter tag can move freely while injecting data. The second is the distance from the backscatter tag to the receiver (receiver distance), which determines the range where the receiver can reliably decode data. The two distances are related to each other because the overall path loss is a combined attenuation of both paths. For the same backscatter communication system, the maximum receiver distance becomes shorter when the transmitter distance increases. However, there are different requirements of the two distances for different applications. For a wearable health sensor that backscatters data to a user's smartphone, e.g., there is no need for a long receiver distance because a phone is presumably always kept nearby. For a sensor deployed on a farm, both the transmitter and the receiver distance need to be long. The different needs of applications provide the possibility to tackle the problem of communication range in different ways.

FM backscatter [15] is an example that makes use of ambient RF signals already having good strength and coverage to solve the problem of transmitter distance. In this system, ambient FM radio is leveraged as the signal source, and the receiver is a commodity FM radio device. This setup enables backscatter tags to be deployed in outdoor environments where it is not possible to set up a dedicated transmitter. It also allows the tags to move freely within a city-scale area without worrying about getting too far from the transmitter. The key contribution of this project is that the backscatter tag can add an FM stream on top of incoming FM-modulated audio. It is achieved by using the frequency modification technique explained in the "Tutorial on Backscatter Communication" section. The backscatter tag shifts the frequency of incoming signals according to the data to be injected, which essentially performs FM on top of the existing FM signal. The receiver demodulates the signal as normal FM and gets the addition of the original FM audio and the injected data. Three operation modes are proposed based on this technique: overlay, where an audio stream is combined with existing radio program; stereo, where the underused stereo band of an FM radio station is used to accommodate reflected signal to eliminate interference to other radio stations; and cooperative, where two receivers work together to remove the original FM audio. This system is able to achieve 3.2 kilobit/s at ranges of 1.5–18 m.

The design of LoRa backscatter [14] develops advanced modulation schemes on backscatter tags to support both long transmitter distance and long receiver distance. Specifically, it is the first wide-area backscatter communication system that uses LoRa. LoRa is chosen because it demonstrates excellent sensitivity at the receiver of −149 dBm, which is key to battle significant path loss as well as attenuation at the backscatter tag. Also, it is more robust to out-of-band interference, such as the excitation signal from the transmitter. In this system, a dedicated RF power emitter transmits a continuous single-tone signal. The signal is modified into a LoRa signal by the backscatter tag and is then received by a commodity LoRa receiver device. However, it is nontrivial to use backscatter tags to generate CSS-modulated LoRa signals. Because CSS uses chirps that linearly increase in frequency, the backscatter tag has to continuously and smoothly change the frequency of the reflected signal. It could be difficult to implement such operation on backscatter tags, because the control logic is all digital. The key innovation here is a new backscatter tag design that combines analog and digital circuits to be able to perform CSS. The design of the backscatter tag is shown in Figure 6. A low-power baseband processor outputs a continuously increasing digital signal, and it is then converted to increasing voltage by a DAC. The voltage is then fed into a VCO and generates a square wave signal whose frequency continuously increases. It is then used to generate a CSS-modulated signal by the RF switch. To avoid interference between the original signal and the backscattered signal, the backscatter tag uses the frequency modification technique mentioned in the "Tutorial on Backscatter Communication" section to move the reflected signal off the original channel. In addition, this work proposes a way to cancel harmonics generated when reflecting the signal to prevent the neighboring channels from being interfered with by the backscatter tag. This is done by using a single-pole multithrow RF switch that has multiple states instead of two, so that the signal multiplied with the incoming signal is more like a sine wave, resulting in weaker sidebands. This technology achieves a range of 475 m.

LoRea [11] is another backscatter system that achieves both long transmitter and receiver distance. In this system, a Wi-Fi or IEEE 802.15.4 chip is put into a special test mode to generate a single-tone excitation signal at 2.4 GHz and 868 MHz, respectively. The receiver and backscatter tag are customized hardware. The system operates at a low bit rate of 2.9 kilobit/s to allow the use of ultrasensitive narrow-band receivers. The backscatter tag modulates data using OOK or FSK. To avoid self-interference, the backscatter tag shifts the reflected signal away from the excitation signal. A highlight of the tag design is that it uses an oscillator instead of an MCU or FPGA to generate the square wave signal used for FS and FSK. This saves power and simplifies the design of the tag. When the transmitter distance is 1 m, it can achieve 3.4 km of receiver range when operating at 868 MHz and 225 m at 2.4 GHz.

The communication range is highly relevant to receiver sensitivity and, thus, is often traded for bit rate. LoRa backscatter uses the modulation scheme of LoRa, e.g., and therefore achieves a data rate up to only 37.5 kilobit/s. Most other backscatter communication systems achieving a much higher throughput with other modulation schemes often come with a poorer sensitivity.

Deployment cost

A typical backscatter communication system includes a backscatter tag and (carrier) transmitter and receiver as supporting devices. While the backscatter tag is usually cheap and tiny, the transmitter and the receiver are often expensive and bulky, greatly increasing the cost of deployment. Unlike conventional wireless communication, such as Wi-Fi and Bluetooth, backscatter communication currently has a much smaller market, and, hence, there is little incentive for manufacturers to massively produce the supporting devices at low cost. With RFID, the most mature backscatter communication technology, e.g., a typical ultrahigh-frequency RFID reader weighs about 0.5 kg [18] and costs more than US$500, which is often beyond the space and cost budget of many IoT deployments, such as smart home applications. Most recent backscatter technology innovations rely on professional equipment like software-defined radio devices or even their own customized hardware, which can potentially discourage customers from using them due to the initial investment, especially when there is already infrastructure for conventional wireless communication like LTE and Wi-Fi. As a result, the high deployment cost challenges the practical adoption of backscatter communication.

Because building customized transmitter and receiver hardware for backscatter communication from scratch would be expensive, it is desirable to leverage the cheaper and common commodity devices and add to them the functionality as a backscatter transmitter or receiver. Such commodity devices could be conventional wireless devices, such as Wi-Fi APs or Bluetooth-enabled computers. Those devices are usually affordable and are already in many scenarios where backscatter communication is to be deployed, so making use of them can greatly reduce the deployment overhead. In addition, commodity wireless devices usually use mature technologies, and their performance has been optimized by significant engineering efforts over the years. Many Wi-Fi chips, e.g., provide excellent sensitivity as good as −80 dBm while costing only several dollars. Hence, making use of these products has a side advantage of improving the system performance.

To achieve the goal of using commodity hardware as backscatter transmitter/receiver, two main problems have to be solved. First, the backscatter tag has to be able to effectively inject data on an excitation signal transmitted by commodity devices, such as Wi-Fi or Bluetooth packets. This could be difficult because backscatter tags do not have the ability to decode the signal and must inject data blindly. Second, the commodity receiver has to be able to decode and process a signal that is modified and reflected by a backscatter tag. The challenge lies in the fact that commodity wireless devices usually allow little control over the decoding process. Packets modified by a backscatter tag may get ignored or corrupted during decoding and never reach a user-space program.

Wi-Fi backscatter [8] presents a way to inject data by modifying the received signal strength indicator (RSSI) or channel state information (CSI) of the Wi-Fi. Almost all Wi-Fi chips provide RSSI information, and many of them also provide CSI, which is a set of complex numbers representing the change of amplitude and phase of each subcarrier. In this system, a commodity Wi-Fi AP is used as the transmitter, and an Intel 5300 wireless network interface controller is used as the receiver. Normal Wi-Fi packets are being transmitted continuously from the transmitter to the receiver. A typical backscatter tag uses an RF switch to change the impedance of the antenna. When set at different states, the backscatter tag affects the Wi-Fi channel between transmitter and receiver differently, resulting in different RSSI and CSI captured at the receiver, used to represent 0 and 1. On the receiver, signal processing is used to reliably detect changes in RSSI and CSI to recover data injected by the backscatter tag. To avoid disrupting decoding packets at the receiver, the RF switch stays in the same state over the duration of multiple Wi-Fi packets. The system achieves up to 1 kilobit/s of throughput with a range of up to 2.1 m. When it is hard to access low-level information, such as RSSI or CSI, an alternative approach is proposed in FS backscatter [19]. Instead of operating on the same channel as the transmitter, the receiver listens to an adjacent channel. The backscatter tag is able to shift incoming Wi-Fi or Bluetooth packets to the adjacent channel. By toggling between shifting and not shifting, the tag is able to transmit data. FS backscatter is able to achieve up to 4.8 m of communication range.

While Wi-Fi backscatter can use commodity Wi-Fi devices as the receiver, it is not compatible with the Wi-Fi protocol itself. There are projects that are compatible with major wireless communication protocols, and both transmitter and receiver can be replaced by commodity devices. The design of intertechnology backscatter [12] proposes a way to transform a valid Bluetooth signal to valid a Wi-Fi or Zigbee signal on a backscatter tag so that a commodity Bluetooth radio can be used as the transmitter and a commodity Wi-Fi or Zigbee radio can be used as the receiver. The key idea behind this system is that, with a carefully designed payload, a commodity Bluetooth radio can transmit a single-tone signal, which can then be modulated into a valid Wi-Fi or Zigbee signal using the methods explained in the "Tutorial on Backscatter Communication" section. This is possible because Bluetooth uses Gaussian FSK, so a continuous stream of 0 s or 1 s is modulated into a single-frequency tone. To achieve the goal of modulating continuous 0 s or 1 s, however, the payload has to be carefully designed because Bluetooth uses a linear feedback shift register to perform data whitening before modulation. This process is inverted to get a payload that will result in all 0 s or 1 s after whitening. In addition, an envelope detector is used on the backscatter tag to detect the start of a Bluetooth packet and help skip the metadata part, because this part is not a single-tone signal. On the backscatter tag, the data of a valid Wi-Fi packet are generated. Methods mentioned in the "Tutorial on Backscatter Communication" section are used to perform DBPSK and DQPSK, two modulation schemes used by Wi-Fi. The tag shifts the signal from the Bluetooth channel to the Wi-Fi channel. The same procedure can also be applied to backscatter Zigbee signals from Bluetooth.

HitchHike [10] presents another project that is compatible with a commodity wireless protocol. This design proposes a way to enable backscatter tags to transform a valid Wi-Fi 802.11b packet to another valid one while modifying the content to inject data. In this system, two commodity Wi-Fi radios act as the backscatter transmitter and receiver. The key innovation is a method called *codeword translation* by which the backscatter tag can transform a valid 802.11b codeword to another valid one so that it can modify some bits in a packet while still allowing the receiver to decode the modified packet. The 0 s and 1 s are represented as a transformed codeword and an untransformed codeword, respectively. The receiver then compares the modified packet with the original one and performs an exclusive operation to recover the data injected by the backscatter tag. Codeword translation is implemented with the RF switch on the backscatter tag. Wi-Fi 802.11b 1 megabit/s uses DBPSK modulation, and there are only two valid codewords representing 0 and 1, with one codeword being the other flipped. To transform a codeword into the other, the backscatter tag needs to flip the phase of the incoming signal, which can be done by using the phase modification method explained earlier. To avoid the backscatter signal and the original signal interfering with each other, the backscatter also shifts the signal to another valid Wi-Fi channel. The system can reach a throughput of up to 300 kilobit/s at 34 m. The idea of codeword translation can also be applied to other wireless protocols, such as Zigbee and Bluetooth [17].

For those projects that are not compatible with existing technologies, they trade deployment cost for other purposes. BackFi [6], e.g., uses a customized transmitter/receiver device that supports concurrent transmission of the transmitter and the backscatter tag. This improves the spectrum efficiency but requires new hardware. LoRa backscatter [14] requires a dedicated transmitter to provide a single-tone excitation signal. This enables the backscatter tag to synthesize a CSS-modulated signal, which significantly improves the communication range. Battery-free cell phones [13] require a special base station to transmit an AM voice signal to the backscatter tag and to receive an FM voice signal from the backscatter tag, and it consumes only 3.48 μW in operation.

Applications empowered by backscatter communication

We envision the prevalence of backscatter tags featured with ultralow power or even battery free can greatly mitigate or even eliminate the existing deployment hurdles of many IoT applications, such as universal localization, ubiquitous surveillance, and invasive monitoring. In the following, we elaborate on how these applications can benefit from backscatter communication along with the research challenges to be addressed.

Universal localization

Location is becoming a fundamental service in mobile/IoT sensing. The tracking demand is extending from smartphones and wearables to universal objects, such as wallets, keys, and pill bottles. Previous research efforts can be classified into two categories: in active approaches, the object of interest needs to emit signals at the milliwatt level and, thus, often carry a battery, whereas passive methods, such as RFID, often require dedicated and costly reader deployment. With backscatter communication, it is possible to attach such battery-free tags to any objects and make them work with existing infrastructure. However, localizing a backscatter tag is not that straightforward, and a number of challenges need to be addressed, such as the RSSI or CSI obtained at the receiver (e.g., AP) being dependent on the location of both tag and transmitter. WiTag [20] presents the first design to achieve that goal—it estimates the angle of arrival from the tag to multiple APs and uses triangulation techniques to achieve 0.92- and 1.48-m median localization error in line-of-sight and nonline-of-sight deployments, respectively, in an office building with commodity Wi-Fi APs.

Ubiquitous surveillance

Wireless cameras are increasingly important and popular for security purposes in the home and office and for public safety. Unfortunately, they still need to be externally powered by outlets, which prevents them from reaching inaccessible areas, such as fabrication plants. By cutting both Internet and power cords, the deployment scale of wireless cameras can potentially reach a new milestone. However, a number of challenges need to be addressed when building a practical video surveillance system based on backscatter communication. The most obvious one can be the mismatch between the intermittent kilobits per second the state-of-the-art solutions can provide and the megabits per second streaming requirement for the application. Performing conventional codec and compression is also extremely challenging on backscatter tags with impoverished compute power. The design of WISPCam [21] opens the door in this direction. It features a battery-free camera that is able to emit a new 176 × 144 gray-scale picture captured and transmitted approximately every 15 min when it is placed 5 m away from a normal RFID reader. The most recent analog video backscatter design [22] even supports 720-p full-high-definition video streaming at 10 frames per second up to distances of 4.9 m from the reader.

Invasive monitoring

In the scenarios where sensor devices need to be instrumented in an invasive manner, the backscatter communication can be found particularly useful because it can potentially eliminate the need for replacing batteries and the accompanying extra high cost in certain applications, such as structural health monitoring (SHM) [23] and implantable health-care monitoring (IHM) [24]. To be more specific, periodic and effective SHM is vital to ensuring safe and reliable operation of large-scale structures (e.g., railways, pipelines, dams, bridges, and aircraft). Deterioration (such as corrosion and fatigue) and damage can be detected at an early stage, and action can be taken correspondingly. The huge labor cost and potential safety issues induced by human inspection today can be addressed with battery-free IoT solutions with backscatter communication. In biomedical applications, IHM poses several harsh requests in the design of implanted medical devices. They need to be tiny and long lasting and radiate low heat. Backscatter again serves as an ideal solution to fulfill all of these requirements. Nevertheless, fundamental tradeoffs existing in power consumption, communication range, bit rate, and form factor need to be fully considered and respected by experts from different domains when designing a backscatter solution dedicated for each use case.

Open areas and future directions

Advanced modulation scheme

Supporting an advanced modulation scheme (on backscatter link) is one of the keys for achieving a higher data rate. The idea of OFDM-based backscatter communication, e.g., has been exercised in both simulation [25] and implementation [17]. However, their throughput is constrained by the fact that OFDM uses much longer symbols. Specifically, the maximum throughput of FreeRider is 60 kilobit/s, in comparison to 300 kilobit/s of HitchHike [10], which uses BPSK and shorter symbols. Presently, there is no efficient design for that in the context of backscatter communication. It would be desirable if new techniques were developed for backscatter tags to modify or generate OFDM signals not only in a low-power manner but also to provide a higher data rate.

Downlink and full duplex

The current backscatter downlink design leverages the low-power RF envelope detector and uses the presence and length of the excitation signal to demodulate data, which are intrinsically low in rate because of the modulation scheme. However, the low-power requirement also poses challenges to implement complex digital signal processing on the tag and, hence, an efficient downlink solution. It would be desirable for the tag's receiver to be renovated to support both efficient and low-power modulation. However, given the limited throughput of the current downlink designs, it would be more efficient if the tag could transmit and receive at the same time (i.e., full duplex) to improve the overall throughput. The key challenge is that the uplink requires the persistent excitation signal, whereas the downlink leverages the intermittent patterns. Obviously, there is a fundamental tradeoff here. Prior work [26] has demonstrated a full-duplex design that achieves 1 kilobit/s downlink and 100 bits/s uplink between two tags, which opens the door for more efforts in this direction.

Multiple-input, multiple-output

As another essential technology extensively used in today's wireless systems for performance enhancement, multiple-input, multiple-output (MIMO) is demanding to be introduced into backscatter system design from different aspects. Beamforming, e.g., can effectively help the backscatter tag to get a strong excitation signal. As another example, the diversity technique can be used potentially in a distributed manner for improving the bit rate and robustness of the tag-to-receiver link. While the idea of MIMO backscatter has been explored in an analytical model [27], there has yet to be real-world implementation to demonstrate its practicality.

Multiple access

It is important to efficiently support multiple access as the backscatter communication technique scales to the network level. Recent work has shown that parallel decoding is an effective physical layer approach. Laissez-faire [28] demonstrated support of an aggregated throughput of 100 kilobit/s for up to 16 devices using parallel decoding, and FlipTracer [29] supports 500 kilobit/s for five tags. In the link layer, FreeRider [17] provides an aggregated throughput of 15 kilobit/s for 20 devices by implementing a basic media access control layer. However, there is yet to be a more efficient design to support large-scale deployment.

Alternative communication medium

Although the performance of radio communication is often limited by scarce spectrum resources, visible light communication (VLC) is always regarded as a complementary solution because it features sufficient spectrum and directionality and is sniff-proof. PassiveVLC [30] presents the state-of-the-art design achieving 1 kilobit/s by modulating the light retroreflection with a commercial liquid-crystal display shutter; yet it also needs to address the challenges for a higher rate, a longer range, and multiple access.

Acknowledgments

This work is supported in part by the National Key Research and Development Plan, China (grant 2016YFB1001200), the Science and Technology Innovation Project of Foshan City, China (grant 2015IT100095), and the Science and Technology Planning Project of Guangdong Province, China (grant 2016B010108002).

Authors

Chenren Xu (chenren@pku.edu.cn) received his B.E. degree in automation from Shanghai University, China, and his Ph.D. degree in computer engineering from Rutgers University, New Brunswick, New Jersey, in 2014. He has been an assistant professor in the Department of Computer Science and a member of the Center for Energy-Efficient Computing and Applications at Peking University, Beijing, China, since 2015. He has held a postdoctoral fellowship at Carnegie Mellon University, Pittsburgh, Pennsylvania, and visiting positions at AT&T Shannon Labs, Florham Park, New Jersey, and Microsoft Research Asia, Beijing, China. His research interests lie in the intersection of wireless, systems, and networking, with a current focus on high-mobility data networking, visible light backscatter communication, and wireless edge systems.

Lei Yang (leiy@mit.edu) received his B.S. degree in computer science from Peking University, Beijing, China, in 2018. He is currently a Ph.D. student in the Networks and Mobile Systems Group, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge. His research interests include computer networks, networked systems, and mobile computing.

Pengyu Zhang (pyzhang@cs.stanford.edu) received his B.E. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, in 2007 and 2010, respectively, and his Ph.D. in computer science from the University of Massachusetts, Amherst in 2016. He has been a postdoctoral researcher with Prof. Sachin Katti at Stanford University, California, since 2016. His research interests include embedded systems, the Internet of Things, cyberphysical systems, and wireless systems. He is a recipient of the Association for Computing Machinery's Special Interest Group on Mobility of Systems, Users, Data and Computing Doctoral Dissertation Award.

References

[1] R. Want, "An introduction to RFID technology," *Pervasive Comput.*, vol. 5, no. 1, pp. 25–33, 2006.

[2] S. Sudevalayam and P. Kulkarni, "Energy harvesting sensor nodes: Survey and implications," *Commun. Surveys Tuts.*, vol. 13, no. 3, pp. 443–461, 2011.

[3] M.-L. Ku, W. Li, Y. Chen, and K. J. Ray Liu, "Advances in energy harvesting communications: Past, present, and future challenges," *Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1384–1412, 2016.

[4] P. Hu, P. Zhang, M. Rostami, and D. Ganesan, "Braidio: An integrated active–passive radio for mobile devices with asymmetric energy budgets," in *Proc. ACM Special Interest Group Data Communication (SIGCOMM)*, 2016, pp. 384–397.

[5] V. Talla and J. R. Smith, "Hybrid analog–digital backscatter: A new approach for battery-free sensing," in *Proc. Annu. Conf. RFID*, 2013, pp. 74–81.

[6] D. Bharadia, K. R. Joshi, M. Kotaru, and S. Katti, "BackFi: High throughput WiFi backscatter," in *Proc. ACM Special Interest Group Data Communication (SIGCOMM)*, 2015, pp. 283–296.

[7] V. Liu, A. Parks, V. Talla, S. Gollakota, D. Wetherall, and J. R. Smith, "Ambient backscatter: Wireless communication out of thin air," in *Proc. ACM Special Interest Group Data Communication (SIGCOMM)*, 2013, pp. 39–50.

[8] B. Kellogg, A. Parks, S. Gollakota, J. R. Smith, and D. Wetherall, "Wi-Fi backscatter: Internet connectivity for RF-powered devices," in *Proc. ACM Special Interest Group Data Communication (SIGCOMM)*, 2014, pp. 607–618.

[9] B. Kellogg, V. Talla, S. Gollakota, and J. R. Smith, "Passive Wi-Fi: Bringing low power to Wi-Fi transmissions," in *Proc. USENIX Symp. Networked Systems Design and Implementation (NSDI)*, 2016, pp. 151–164.

[10] P. Zhang, D. Bharadia, K. R. Joshi, and S. Katti, "HitchHike: Practical backscatter using commodity WiFi," in *Proc. ACM Conf. Embedded Networked Sensor Systems (SenSys)*, 2016, pp. 259–271.

[11] A. Varshney, O. Harms, C.-P. Penichet, C. Rohner, F. Hermans, and T. Voigt, "LoREA: A backscatter architecture that achieves a long communication range," in *Proc. ACM Conf. Embedded Networked Sensor Systems (SenSys)*, 2017, pp. 18:1–18:14.

[12] V. Iyer, V. Talla, B. Kellogg, S. Gollakota, and J. Smith, "Inter-technology backscatter: Towards Internet connectivity for implanted devices," in *Proc. ACM Special Interest Group Data Communication (SIGCOMM)*, 2016, pp. 356–369.

[13] V. Talla, B. Kellogg, S. Gollakota, and J. R. Smith, "Battery-free cellphone," in *Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Computing (UBICOMP)*, 2017, pp. 25:1–25:20.

[14] V. Talla, M. Hessar, B. Kellogg, A. Najafi, J. R. Smith, and S. Gollakota, "LoRA backscatter: Enabling the vision of ubiquitous connectivity," in *Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Compting (UBICOMP)*, 2017, pp. 105:1–105:24.

[15] A. Wang, V. Iyer, V. Talla, J. R. Smith, and S. Gollakota, "FM backscatter: Enabling connected cities and smart fabrics," in *Proc. USENIX Symp. Networked Systems Design and Implementation (NSDI)*, 2017, pp. 243–258.

[16] C. Mikeka and H. Arai, "Design issues in radio frequency energy harvesting system," in *Sustainable Energy Harvesting Technologies-Past, Present and Future*, Y. K. Tan, Ed. London: InTech Open, 2011, pp. 235–256.

[17] P. Zhang, C. Josephson, D. Bharadia, and S. Katti, "Freerider: Backscatter communication using commodity radios," in *Proc. Int. Conf. Emerging Networking Experiments and Technologies (CoNEXT)*, 2017, pp. 389–401.

[18] Y. Ma, N. Selby, and F. Adib, "Drone relays for battery-free networks," in *Proc. ACM Special Interest Group on Data Communication (SIGCOMM)*, 2017, pp. 335–347.

[19] P. Zhang, M. Rostami, P. Hu, and D. Ganesan, "Enabling practical backscatter communication for on-body sensors," in *Proc. ACM Special Interest Group Data Communication (SIGCOMM)*, 2016, pp. 370–383.

[20] M. Kotaru, P. Zhang, and S. Katti, "Localizing low-power backscatter tags using commodity WiFi," in *Proc. Int. Conf. Emerging Networking Experiments and Technologies (CoNEXT)*, 2017, pp. 251–262.

[21] S. Naderiparizi, A. N. Parks, Z. Kapetanovic, B. Ransford, and J. R. Smith, "WISPCam: A battery-free RFID camera," in *Proc. Annu. Conf. RFID*, 2015, pp. 166–173.

[22] S. Naderiparizi, M. Hessar, V. Talla, S. Gollakota, and J. R. Smith, "Towards battery-free HD video streaming," in *Proc. USENIX Symp. Networked Systems Design and Implementation (NSDI)*, 2018, pp. 233–247.

[23] J. Zhang, G. Yun Tian, A. M. J. Marindra, A. Imam Sunny, and A. B. Zhao, "A review of passive RFID tag antenna-based sensors and systems for structural health monitoring applications," *Sensors*, vol. 17, no. 2, pp. 265:1–265:33, 2017.

[24] A. Darwish and A. E. Hassanien, "Wearable and implantable wireless sensor network solutions for healthcare monitoring," *Sensors*, vol. 11, no. 6, pp. 5561–5595, 2011.

[25] D. Darsena, G. Gelli, and F. Verde, "Modeling and performance analysis of wireless networks with ambient backscatter devices," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1797–1814, 2017.

[26] V. Liu, V. Talla, and S. Gollakota, "Enabling instantaneous feedback with full-duplex backscatter," in *Proc. Annu. Int. Conf. Mobile Computing and Networking*, 2014, pp. 67–78.

[27] C. Boyer and S. Roy, "Backscatter communication and RFID: Coding, energy, and MIMO analysis," *IEEE Trans. Commun.*, vol. 62, no. 3, pp. 770–785, 2014.

[28] P. Hu, P. Zhang, and D. Ganesan, "Laissez-faire: Fully asymmetric backscatter communication," in *Proc. ACM Special Interest Group Data Communication (SIGCOMM)*, 2015, pp. 255–267.

[29] M. Jin, Y. He, X. Meng, Y. Zheng, D. Fang, and X. Chen, "Fliptracer: Practical parallel decoding for backscatter communication," in *Proc. 23rd Annu. Int. Conf. Mobile Computing and Networking*, 2017, pp. 275–287.

[30] X. Xu, Y. Shen, J. Yang, C. Xu, G. Shen, G. Chen, and Y. Ni, "PassiveVLC: Enabling practical visible light backscatter communication for battery-free IoT applications," in *Proc. 23rd Annu. Int. Conf. Mobile Computing and Networking,* 2017, pp. 180–192.

SP

SIGNAL PROCESSING AND THE INTERNET OF THINGS

Aggelos Bletsas, Panos N. Alevizos,
and Georgios Vougioukas

The Art of Signal Processing in Backscatter Radio for μW (or Less) Internet of Things

Intelligent signal processing and backscatter radio enabling batteryless connectivity

INTERNET OF THINGS—ISTOCKPHOTO.COM/IAREMENKO
CIRCUITS—IMAGE LICENSED BY INGRAM PUBLISHING

Backscatter (or simply *scatter*) radio is based on reflection principles, where each tag modulates information on top of an illuminating signal, by simply connecting its antenna to different loads; modulation of information is based on the modifications of the tag antenna-load reflection coefficient, requiring in principle only a switch and omitting power-consuming signal conditioning units, such as mixers, amplifiers, oscillators, and filters. The ultralow-power nature of backscatter radio, in conjunction with the recent advances in multiple access and achieved communication ranges (on the order of hundreds of meters to kilometers), due to intelligent signal processing, elevate backscatter radio as the de facto communication principle for μW (or less)-level consumption, last-mile connectivity, and Internet of Things (IoT) networking. This article is an update to the state-of-the-art advances in the emerging backscatter radio domain, focusing on the signal processing engine, including ambient illumination from existing signals, as well as unconventional backscatter radio-based IoT technologies that could revolutionize environmental sensing and agriculture. Finally, the offered research methodology and techniques in short-packet, channel-encoded (or not), coherent (or not) sequence detection will assist researchers in radio-frequency identification (RFID)/backscatter radio as well as other domains of the telecommunications industry.

Introduction

RFID is based on backscatter, i.e., reflection radio, where each tag modulates information on top of an illuminating signal by simply connecting its antenna to different loads (Figure 1); modulation of information is based on the modifications of the tag antenna-load reflection coefficient, requiring in principle only a (transistor) switch and an antenna to reflect information, omitting power-consuming signal conditioning and generating units, such as mixers, amplifiers, oscillators, and filters. Thus, backscatter radio is a promising solution for ultralow-power radio communication and networks [1]. Recent work has demonstrated backscatter communication with a few

Digital Object Identifier 10.1109/MSP.2018.2837678
Date of publication: 28 August 2018

μW consumption at the tag [2]–[4] even at continuous, nondutycycled operation [5].

The basic limitation of passive RFIDs in terms of communication range, on the order of meters, stems from the fact that they are batteryless, harvesting their required energy from the illuminating signal; the bottleneck element has been the RF harvesting circuitry sensitivity and not the backscatter radio principle. Seminal work in [1] was the first to decouple RF harvesting from backscatter radio and showed that semipassive tags, i.e., reflection radio tags with an external power source (e.g., a coin battery) could be received by a software-defined radio (SDR) with extended communication ranges. Work in [1] highlighted the idiosyncracies of backscatter radio, e.g., the fact that tag modulation occurs at passband; thus, orthogonal signaling, e.g., frequency-shift keying (FSK) reception using the detectors for conventional (Marconi) radio, would result in a 3-dB loss, since half of the useful signal (and appropriate matched filters) would be overlooked. Orthogonal switching signaling among multiple tags allowed for collision-free multiple access, even with a common carrier, while noncoherent, symbol-by-symbol detection of continuous phase FSK, i.e., minimum-shift keying (MSK), was demonstrated in SDR.

The need for low-complexity, resource-constrained tags, as well as the basic requirement for extended communication range, coverage, and fast, low-complexity reception, imposes additional challenging requirements in terms of nontrivial signal processing at the reader. Decoupling the illuminating emitter from the receiver of the backscattered signals (bistatic architecture) offers flexibility and better link budgets at the expense of additional channel unknowns, since emitter-to-tag and tag-to-reader links become distinct (see Figure 2). Additionally, the tag-reflected packets must be relatively short to reduce energy consumption at the tag and expedite the processing at the reader in network setups, with multiple tags operating simultaneously.

Reflector/tag: Scatter radio principles

The simplest case of backscatter radio utilizes only two passive loads: the tag/reflector modulates information by modifying the refection coefficient of the tag antenna and connected load. In that way, the induced signal at the tag antenna, stemming from a distant illuminator, is reflected back with modified amplitude and phase.

More specifically, a modified reflection coefficient is defined as $\Gamma_i = (Z_i - Z_a^*)/(Z_i + Z_a)$, where $i \in \{0,1\}$ for the two loads Z_0, Z_1 and Z_a is the (complex in general) tag antenna characteristic impedance at the utilized carrier frequency. The baseband equivalent signal, when the tag antenna is connected at load Z_i, with corresponding reflection coefficient Γ_i, $i \in \{0,1\}$, is given by

$$A_s - \Gamma_i,$$

where A_s is the (complex) load-independent tag antenna structural mode; the latter depends on the geometry and construction materials of the antenna. More than two loads and hence, multiple bits per load, have been also recently demonstrated with energy-efficient circuits [6], [7].

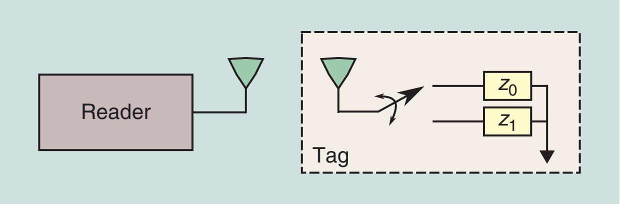


FIGURE 1. The backscatter radio principle: information is modulated on reflection at the tag of an illuminating signal, using (at least) two loads; only switching at the tag between loads is needed, omitting power-consuming signal conditioning and generating modules (e.g., amplifiers).

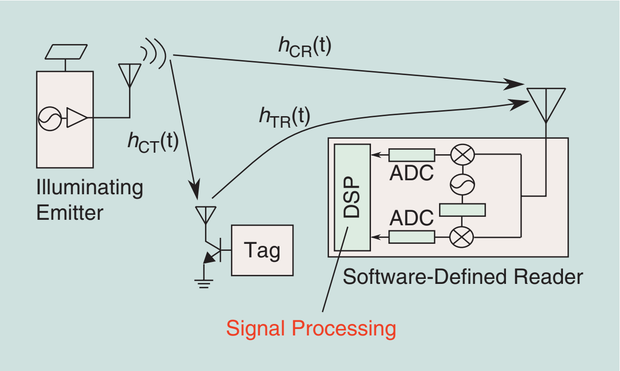


FIGURE 2. The backscatter radio principle: intelligent signal processing at the receiver allows for extended communication ranges and tag networking.

Orthogonal and nonorthogonal tag signaling

The simplest case for binary tag modulation occurs when the tag terminates its antenna at load for the whole bit duration T, i.e., at Z_0 (Γ_0) for bit "0" and Z_1 (Γ_1) for bit "1." This is the case of nonorthogonal signaling, utilized in industrial RFIDs, commonly referred to as *on-off keying* (*OOK*). Thus, assuming that $\Gamma_{\text{tag}} = \Gamma_0$ for $x_n = -1, \Gamma_{\text{tag}} = \Gamma_1$ for $x_n = +1$, where x_n is the (binary) information of the nth bit, the baseband equivalent of the tag-backscattered signal is given by

$$A_s - \Gamma_{\text{tag}} = \left(A_s - \frac{\Gamma_0 + \Gamma_1}{2}\right) + x_n \frac{\Gamma_0 - \Gamma_1}{2}, x_n \in \{\pm 1\}, \quad (1)$$

$$\Rightarrow x_{\text{Tag}}(t) = \left(A_s - \frac{\Gamma_0 + \Gamma_1}{2}\right) + \frac{\Gamma_0 - \Gamma_1}{2} x_n \Pi_T(t - nT),$$
$$t \in [nT, (n+1)T), \quad (2)$$

$\Pi_T(t) = 1$ for $t \in [0, T)$ and zero elsewhere.

Alternatively, the tag can continuously switch between the two loads during bit reflection, with switching frequency F_0 for bit "0" or F_1 for bit "1." This is the case of orthogonal signaling, as in FSK. If the switching pattern of the tag during the nth bit has a fundamental frequency F_0 (period $1/F_0$) for bit "0" ($x_n = -1$) and F_1 for bit "1" ($x_n = 1$), the baseband

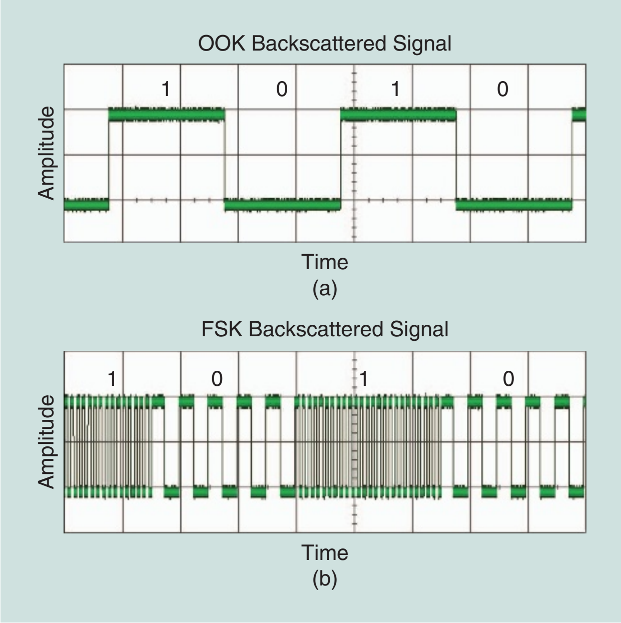


FIGURE 3. The time-domain amplitude of tag-modulated backscatter (complex) baseband signal: (a) OOK and (b) FSK.

equivalent of the tag backscattered signal for the nth binary-modulated information bit is given by

$$x_{\text{Tag}}(t) = \left(A_s - \frac{\Gamma_0 + \Gamma_1}{2}\right) + \frac{\Gamma_0 - \Gamma_1}{2}\, b_n(t - nT),$$
$$t \in [nT, (n+1)T), \tag{3}$$

where $b_n(t)$ is a (periodic) pulse train with fundamental frequency $F_{x_n} \in \{F_0, F_1\}$ and duration equal to bit duration T, with $T \gg \max(1/F_0, 1/F_1)$. Additionally, $|F_0 - F_1| = k/(2T)$ for coherent detection, and $|F_0 - F_1| = k/T$ for noncoherent detection, $k \in \mathbb{Z}$.

For $b_n(t)$ a 50% duty-cycle pulse train and even, i.e., $b_n(t) = b_n(-t)$, the following Fourier series representation holds:

$$b_n(t) = \frac{4}{\pi}\sum_{k=0}^{+\infty}\frac{1}{2k+1}\cos[2\pi(2k+1)F_{x_n}t], \tag{4}$$

i.e., only odd harmonics exist due to 50% duty cycle and only cosines, due to being even. For $b_n(t)$ a 50% duty-cycle pulse train and odd, i.e., $b_n(t) = -b_n(-t)$, the following Fourier series representation holds:

$$b_n(t) = \frac{4}{\pi}\sum_{k=0}^{+\infty}\frac{1}{2k+1}\sin[2\pi(2k+1)F_{x_n}t], \tag{5}$$

i.e., only odd harmonics exist due to 50% duty cycle and only sines, due to being odd. Thus, timing during tag modulation matters, and there is remaining phase Φ at the tag backscattered signal due to imperfect timings during tag modulation, modeled as follows:

$$b_n(t) = \frac{4}{\pi}\sum_{k=0}^{+\infty}\frac{1}{2k+1}\cos[2\pi(2k+1)F_{x_n}t + \Phi]. \tag{6}$$

The amplitude of the baseband tag-backscattered signal for OOK and FSK is shown in Figure 3(a) and (b), respectively, as measured in the laboratory. The advantage of OOK is that it is exploited in Gen2, the industrial RFID protocol, as previously mentioned; the disadvantage is that the spectrum of the tag's backscattered signal is centered at the illuminator's carrier frequency, where extensive reflections from the environment occur, offering clutter noise and limiting signal-to-noise ratio (SNR); furthermore, OOK requires time-domain multiplexing of several tags, requiring a receiver at each tag and carrier-sense multiple access; Gen2 RFIDs utilize framed Aloha. Alternatively, detecting simultaneously backscattering tags can be performed with time-domain, signal-specific techniques at the reader [8].

Figure 4 shows a portion of measured spectrum (at the lab) for binary FSK backscattering, depicting the four peaks of the fundamental frequencies F_0, F_1 around the carrier frequency F_c of the illuminator; there are two peaks for F_0 due to the cosine term of (6) and another two for F_1; it is noted that the peaks due to the (odd) harmonics of (6) are not depicted. FSK is tailored to the power-limited regime, where backscatter operates and allows for receiverless tags, multiplexed at the frequency domain; networking several tags, simultaneously backscattering, becomes easily possible using simple

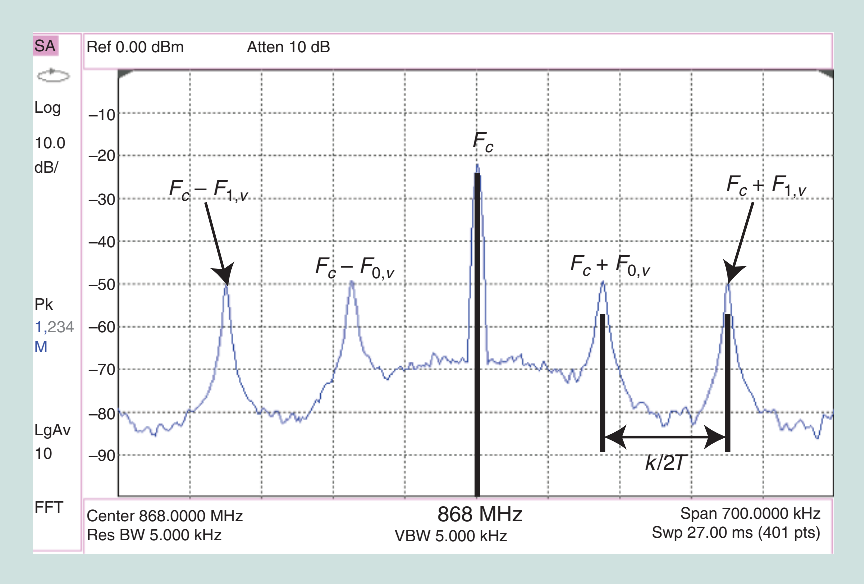


FIGURE 4. Switching between two loads with F_0 (F_1) offers at least two peaks around the illuminator's carrier frequency F_c at $F_c \pm F_0$ ($F_c \pm F_1$). Careful selection of the switching frequencies offers collision-free simultaneous backscattering from multiple tags.

signal processing at the physical layer: all that is required is assigning distinct switching frequency pairs $\{F_{0,\nu}, F_{1,\nu}\}$ for different tags, i.e., $\nu = 1, 2, \ldots, N$ tags [1]. However, such a solution may not be applicable for a large number of high-bit-rate sensors.

This shows a fundamental difference of backscatter radio, compared to conventional Marconi radios: modulation occurs directly at the induced (at the tag) illuminator signal, without any type of upconversion at passband; thus, detection techniques should be tailored to such idiosyncrasy. For example, applying common FSK detection schemes at the signal of Figure 4 could neglect half of the peaks and, thus, half of the useful signal.

Finally, it is noted that there may be a combination of the previously described modulations; each tag can be assigned a unique switching frequency F_ν between the two loads, with a 50% duty cycle, to enjoy frequency-domain multiple access among various tags $\nu = 1, 2, \ldots$ and also exploit modifications of Φ in (6) for the transmission of information [9].

Tag pulse shaping and structural mode

Could only two tag termination loads be used to better shape the backscattered spectrum and improve spectral efficiency? For example, is there any way to alleviate the existence of odd-order harmonics in backscatter FSK, using only two loads at the tag? The answer is given in Lemma 1.

Lemma 1

Pulse shaping in backscatter radio tags with only two loads is possible [1]. Work in [1] utilized two loads and minimum shift keying (MSK), a special variant of binary FSK: instead of (6), where switching frequency changes abruptly at the bit boundaries, the tag implements MSK by continuously changing the instantaneous switching frequency (and, hence, signal phase), so that no discontinuities occur at the bit boundaries; such an operation was performed at the tag using an embedded phase-locked loop, offering power spectral density (PSD) of the backscattered signal that dropped with the fourth power of frequency, as opposed to conventional PAM/quadrature amplitude modulation/PSK (where PSD drops with the square of frequency). MSK can also be seen as offset-quadrature PSK with memory and sinusoid modulating pulses, corroborating its inherent pulse-shaping nature.

Pulse shaping with FSK and more than two loads was recently proposed in [10]: switching alternatively between a series of loads implemented a (rotating) complex phasor, multiplying the induced (at the tag) signal and, thus, shifting its spectrum only right (or left) of the illuminating carrier frequency, depending on the rotating direction. In that way, smaller bandwidth could be utilized.

Finally, it is noted that the current mind-set in backscatter literature dismisses the value of the tag antenna's structural mode. That is due to the fact that, for binary coherent (i.e., minimum distance) detection, the distance between the utilized constellation points $|(A_s - \Gamma_0) - (A_s - \Gamma_1)| = |\Gamma_1 - \Gamma_0|$ matters, which is A_s independent. However, for certain bistatic scenarios (e.g., a blocked illuminator-to-reader link) and certain housekeeping tasks before detection [e.g., carrier frequency offset (CFO) estimation], A_s matters [11]. A measurement and estimation method for A_s can be found in [12].

Universal system model: Monostatic versus bistatic versus ambient

Figure 2 depicts the case of bistatic backscatter radio, where the illuminator of the tag and reader of the tag-backscattered signal are distinct units, placed at different locations. Assuming flat fading, the channel impulse response between illuminator and reader, illuminator and tag, and tag and reader is given by $h_{\mathrm{m}}(t) = a_{\mathrm{m}}\delta(t - \tau_{\mathrm{m}})$, with $\mathrm{m} \in \{\mathrm{CR}, \mathrm{CT}, \mathrm{TR}\}$, respectively; the baseband equivalent for each link is given by $a_m e^{-j2\pi f_c \tau_m}$, where s_f is the utilized carrier frequency. Based on this modeling, the baseband representation of the received signal at the reader is given by (Figure 2)

$$\mathsf{y}(t) = a_{\mathrm{CR}}\mathsf{e}^{-\mathrm{j}\phi_{\mathrm{CR}}}\mathsf{c}(t) + a_{\mathrm{CT}}\mathsf{e}^{-\mathrm{j}\phi_{\mathrm{CT}}}a_{\mathrm{TR}}\mathsf{e}^{-\mathrm{j}\phi_{\mathrm{TR}}}\,\mathsf{s}x_{\mathrm{Tag}}^{\mathrm{mod}}(t) + \mathsf{n}(t), \tag{7}$$

where $\phi_{\mathrm{m}} = 2\pi f_c \tau_{\mathrm{m}}$, with $\mathrm{m} \in \{\mathrm{CR}, \mathrm{CT}, \mathrm{TR}\}$, $\mathsf{c}(t)$ is the signal transmitted by the illuminator, $\mathsf{n}(t)$ models white complex Gaussian noise at the reader, and s models nonidealities in backscattering efficiency at the tag (e.g., due to mismatches, imperfect estimation of load values, etc.). For bistatic setups, the illuminator transmits a simple carrier signal $\mathsf{c}(t) = \sqrt{2P_{\mathrm{C}}}\,\mathsf{e}^{-\mathrm{j}(2\pi\Delta F t + \Delta\phi)}$, with carrier frequency and phase offset between illuminator and reader denoted by ΔF and $\Delta\phi$, respectively, and P_{C} as the illuminator's transmission power; in that case, $x_{\mathrm{Tag}}^{\mathrm{mod}}(t)$ in (7) is given by $x_{\mathrm{Tag}}^{\mathrm{mod}}(t) = \sqrt{2P_{\mathrm{C}}}\,\mathsf{e}^{-\mathrm{j}(2\pi\Delta F t + \Delta\phi)}x_{\mathrm{Tag}}(t)$. Notice that, for monostatic systems where illuminator and reader share the same oscillator, the CFO is zero $(\Delta F = 0)$ and the link carrier reader models the duplexer's imperfection, i.e., the signal leakage from the transmit to the receive chain (e.g., due to circulator's imperfection and coupling effects); the system model (7) can describe both bistatic as well as monostatic setups.

It also noted that the previously described system model can describe asymmetric scenarios, i.e., when the channel statistics between tag and reader are vastly different than the statistics between tag and illuminator; it can also describe the case of ambient illuminators, i.e., when the illuminating signal $\mathsf{c}(t)$ is already modulated; in that case, $\mathsf{c}(t) = \mathsf{m}(t)\mathsf{e}^{-\mathrm{j}(2\pi\Delta F t + \Delta\phi - \phi(t))}$, where $\mathsf{m}(t)\mathsf{e}^{\mathrm{j}\phi(t)}$ is the complex envelope of the ambient illuminator's signal. The previously mentioned bistatic model, where the illuminator was decoupled from the reader, appeared first in [11] and [13]–[15]; work in [16] and [17] studied a similar model, having in mind multiple (colocated) antennas at the reader and multiple antennas at the tag; ambient backscatter [18] is a special case of the bistatic architecture.

Due to the lack of any type of specialized filtering/signal conditioning or amplification, there is no additional noise term induced by the tag at (7). Backscatter communication is power limited, and required signal processing for reliable detection becomes challenging. Furthermore, there are many unknown channel and tag-dependent parameters, pointing toward the direction of noncoherent processing. We will summarize breakthroughs in noncoherent, as well as coherent, processing next. Finally, for ultralow-power and low-bit-rate backscatter sensor networks, short packets must be employed to reduce computation complexity at the tag and computation and decoding complexity at the reader, especially when several streams from multiple tags are processed in parallel. We will also summarize recent findings in short-packet communication techniques that could be beneficial to other domains.

Receiver: Noncoherent processing

Symbol-by-symbol detection

Figure 5 depicts four matched filters, in the form of correlators: two for the signals that correspond to F_0 and another two for the signals that correspond to F_1 (as explained in the previous secton). The correlator is equivalent to a matched filter for perfect synchronization. Before filtering, necessary carrier frequency estimation ($\Delta\hat{F}$) using periodograms and compensation is performed (assuming bistatic setups), as well as dc offset removal (through time average removal); the outcome per bit of such filtering is a 4×1 complex vector $\mathbf{r} = [r_0^+ \; r_0^- \; r_1^+ \; r_1^-]^\top$; work in [11] and [13]–[15] suggested the following energy-based, noncoherent detector:

$$\left|r_0^+\right|^2 + \left|r_0^-\right|^2 \overset{\text{bit }0}{\geq} \left|r_1^+\right|^2 + \left|r_1^-\right|^2. \qquad (8)$$

An immediate question arises: Why is squaring of the amplitudes in (8) required and not simply taking the absolute norm? The answer is provided next.

Subsequent work proved that (8) is the outcome of hybrid composite hypothesis testing (HCHT) symbol-by-symbol detection. Denoting $\mathbb{B} \equiv \{0,1\}, \mathbf{s}_i = [1-i \; 1-i \; i \; i]^\top, i \in \{0,1\}, \Phi_0$ and Φ_1 the value of $\mathbf{\Phi}$ in (6), for bit “0,” bit “1,” respectively, $\mathbf{\Phi} = \{\Phi_0, \Phi_1\}$,

$$\mathrm{E} = \kappa^2 P_{\mathrm{C}} \left|\Gamma_0 - \Gamma_1\right|^2 \mathrm{s}^2 T, h = a_{\mathrm{CT}} \, a_{\mathrm{TR}} \, e^{\mathrm{j}(\phi_{\mathrm{CT}} + \phi_{\mathrm{TR}} + \Delta\phi + \angle(\Gamma_0 - \Gamma_1))}, \qquad (9)$$

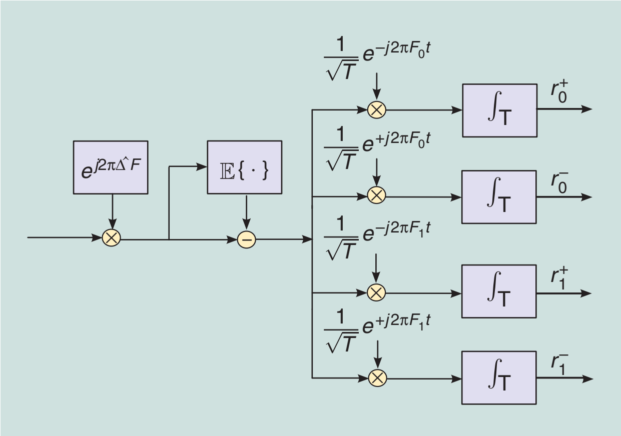

$\frac{1}{\sqrt{T}} e^{-j2\pi F_0 t}$
$\frac{1}{\sqrt{T}} e^{+j2\pi F_0 t}$
$\frac{1}{\sqrt{T}} e^{-j2\pi F_1 t}$
$\frac{1}{\sqrt{T}} e^{+j2\pi F_1 t}$
$e^{j2\pi\Delta\hat{F}}$
$\mathbb{E}\{\cdot\}$
$\int_T$
r_0^+
r_0^-
r_1^+
r_1^-

FIGURE 5. Backscatter FSK, as described by (6) and (7), requires four matched filters, not two.

and $\mathrm{x}_i(\mathbf{\Phi}) = \sqrt{\frac{\mathrm{E}}{2}} \left[\mathrm{e}^{+\mathrm{j}\Phi_0}, \mathrm{e}^{-\mathrm{j}\Phi_0}, \mathrm{e}^{+\mathrm{j}\Phi_1}, \mathrm{e}^{-\mathrm{j}\Phi_1}\right]^\top \odot \mathbf{s}_i, i \in \mathbb{B}$, where $\odot$ denotes point-wise multiplication and κ is a constant, Theorem 1 is presented, assuming complex white Gaussian noise at the receiver [19], [20].

Theorem 1: Noncoherent HCHT symbol-by-symbol backscatter FSK detection

$$\begin{aligned} &\arg\max_{i \in \mathbb{B}} \left\{ \mathbb{E}_{\mathbf{\Phi}} \left[\max_{h \in \mathbb{C}} \ln[\mathrm{f}(\mathbf{r} \,|\, i, h, \mathbf{\Phi})] \right] \right\} \\ &\Leftrightarrow \left|r_0^+\right|^2 + \left|r_0^-\right|^2 \underset{i=1}{\overset{i=0}{\gtrless}} \left|r_1^+\right|^2 + \left|r_1^-\right|^2, \end{aligned} \qquad (10)$$

where $\mathrm{f}(\cdot|\cdot)$ denotes the conditional pdf; the expectation operation (10) gets rid of the unknown phases $\mathbf{\Phi}$, while the maximization operator offers estimation of the unknown channel (and, hence, the hybrid nature of the detector). Interestingly, a pure maximization operation for both unknowns, i.e., a generalized likelihood ratio test (GLRT) receiver, offers the result given in Theorem 2 [19], [20].

Theorem 2: Noncoherent GLRT symbol-by-symbol backscatter FSK detection

$$\begin{aligned} &\arg\max_{i \in \mathbb{B}} \left\{ \max_{\mathbf{\Phi} \in [0,2\pi)^2} \max_{h \in \mathbb{C}} \ln[\mathrm{f}(\mathbf{r} \,|\, i, h, \mathbf{\Phi})] \right\} \\ &\Leftrightarrow \left|r_0^+\right| + \left|r_0^-\right| \underset{i=1}{\overset{i=0}{\gtrless}} \left|r_1^+\right| + \left|r_1^-\right|. \end{aligned} \qquad (11)$$

Work in [1] offered noncoherent symbol-by-symbol detection for backscatter MSK, while work in [11] and [13]–[15] also studied the case of noncoherent symbol-by-symbol for OOK, using energy-based sufficient statistics compared to carefully selected thresholds.

Short-packet/sequence detection—no channel coding

For relatively static environments, channel coherence time can be greater than packet duration, especially when packets are relatively short. Under this assumption, and denoting tag (reflected) information sequence as $\mathbf{i} = [i_1 \, i_2 \ldots i_{N_\mathrm{s}}]^\top \in \mathbb{B}^{N_\mathrm{s}}$, with $N_\mathrm{s} \leq N_{\mathrm{coh}}$ where $N_{\mathrm{coh}} \in \mathbb{N}$ the channel coherence time measured in number of bit periods, and reader received sequence as $\mathbf{r}_{1:N_\mathrm{s}}$, the GLRT sequence detector, assuming complex Gaussian noise at the receiver, is given as follows [21], [22]:

$$\mathbf{i}_{\mathrm{GLRT}} = \arg\max_{\mathbf{i} \in \mathbb{B}^{N_\mathrm{s}}} \max_{\mathbf{\Phi} \in [0, 2\pi)^2} \max_{h \in \mathbb{C}} \ln[\mathrm{f}(\mathbf{r}_{1:N_\mathrm{s}} \,|\, \mathbf{i}, h, \mathbf{\Phi})], \qquad (12)$$

where $\mathrm{f}(\cdot|\cdot)$ denotes the conditional pdf; the above detector, implemented through exhaustive search, requires assessing 2^{N_s} possible sequences; such search, even for moderate sequence length N_s is prohibitive. Fortunately, Theorem 3 unlocks the GLRT potential [20], [21], [22].

Theorem 3

There exists algorithm that finds $\mathbf{i}_{\mathrm{GLRT}}$ with complexity $\mathcal{O}(N_\mathrm{s} \log N_\mathrm{s})$, instead of $\mathcal{O}(2^{N_\mathrm{s}})$. The algorithm provided in [20]–[22] can be applied to any orthogonal signaling, including FSK for backscatter, as well as Marconi radios; given that orthogonal signaling is tailored to the power-limited regime, the applicability of Theorem 3 is wide, for various scenarios in flat fading, terrestrial, underwater, or satellite communications, with channel unchanged during packet/sequence transmission.

For relatively static environments, channel coherence time can be greater than packet duration, especially when packets are relatively short.

FM0 line coding, utilized in industrial (Gen2) RFID, can be seen as orthogonal signaling despite the fact that industrial RFIDs utilize OOK and not FSK. Such interpretation is possible, by observing half-bit before and half-bit after the OOK-modulated/FM0-encoded bit of interest (totaling $2T$ interval for bit duration of T). Thus, Theorem 3 offers noncoherent sequence detection of Gen2/FM0 RFID tags with loglinear complexity in the sequence length and GLRT performance, without utilizing any type of preambles.

Short-packet/sequence detection with channel coding

Relaxing the small delay requirement, interleaving of depth D can be exploited to diminish long bursts of fading while offering reliable communication in a noncoherent fashion. Assuming N_c coded bits per tag-backscattered sequence, interleaving of depth D means that the tag buffers exactly D coded sequences (of length N_c each) and backscatters them column-wise; the reader performs buffering and performs the reverse operation. Following the equivalent FSK signal model found in [23], work in [19] and [20] showed that interleaving for backscatter FSK offers the following:

$$\mathbf{r}_{1:N_c} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_{N_c} \end{bmatrix} = \begin{bmatrix} h_1 \mathbf{x}_{c_1}(\Phi) \\ h_2 \mathbf{x}_{c_2}(\Phi) \\ \vdots \\ h_{N_c} \mathbf{x}_{c_{N_c}}(\Phi) \end{bmatrix} + \begin{bmatrix} \mathbf{n}_1 \\ \mathbf{n}_2 \\ \vdots \\ \mathbf{n}_{N_c} \end{bmatrix}, \tag{13}$$

where $\mathbf{n}_i, i \in \{1, 2, \ldots, N_c\}$ is a 4×1 complex, circularly symmetric Gaussian vector (with such vectors independent for distinct indexes) and the rest of notation is as in the previous sections. For $DT \geq T_{\mathrm{coh}}$, interleaving is comparable to compound channel coefficients $\{h_i\}$ being independent for different $i \in \{1, 2, \ldots, N_c\}$. Thus, the backscattered bits in a specific coded sequence/short packet enjoy statistically independent fading coefficients. Theorem 4 offers soft-decision metrics for noncoherent, channel-coded sequence detection in a structured way [19], [20].

Theorem 4

For $DT \geq T_{\mathrm{coh}}$, noncoherent HCHT soft-decision decoding for channel-coded backscatter FSK amounts to

$$\arg\max_{\mathbf{c} \in C} \left\{ \mathbb{E}_{\boldsymbol{\Phi}} \left[\max_{\mathbf{h} \in \mathbb{C}^{N_c}} \ln [\mathsf{f}(\mathbf{r}_{1:N_c} \mid \mathbf{c}, \mathbf{h}, \boldsymbol{\Phi})] \right] \right\} \Leftrightarrow \arg\max_{\mathbf{c} \in C} \sum_{n=1}^{N_c} w_n c_n, \tag{14}$$

where $w_n \triangleq \left| r_1^+(n) \right|^2 + \left| r_1^-(n) \right|^2 - \left(\left| r_0^+(n) \right|^2 + \left| r_0^-(n) \right|^2 \right), n = 1, 2, \ldots, N_c$, $\mathbf{h} = [h_1\ h_2 \ldots h_{N_c}]$, and $c_n \in \{0, 1\}$.

This allows simple calculation of the most appropriate sequence among all possible coded sequences (denoted as set C). For example, there are 2^{16} possible coded sequences for a 1/2-rate code with sequence length $N_c = 32$. For small packets/coded sequences, as targeted in this work, such exhaustive search with the aforementioned weights is feasible. Other selection of weights is also possible [24]. Experimental results with Bose–Chaudhuri–Hocquenghem (BCH) and Reed–Muller (RM) codes, ultralow-cost, 8-b, microcontroller-based tags and an SDR-based reader can be found in [19], [20], and [24].

Receiver: Coherent processing

Short-packet/sequence detection with/without coding

Work in [23] offered a simplified baseband signal representation for backscatter FSK per bit [23, Th. 1]:

$$\mathbf{r} = [r_0^+\ r_0^-\ r_1^+\ r_1^-]^\top = h\sqrt{\frac{\mathrm{E}}{2}} [\mathrm{e}^{+\mathrm{j}\Phi_0}\ \mathrm{e}^{-\mathrm{j}\Phi_0}\ \mathrm{e}^{+\mathrm{j}\Phi_1}\ \mathrm{e}^{-\mathrm{j}\Phi_1}]^\top \odot \mathbf{s}_i + \mathbf{n}, \tag{15}$$

where the notation follows as in the section "Symbol-by-Symbol Detection," e.g., $\mathbf{s}_i = [1 - i\ \ 1 - i\ i\ i]^\top$, $i \in \{0, 1\}$.

When the tag reflects a known (to the reader) preamble, the reader can find out the estimate

$$\hat{\mathbf{h}} = [\hat{h}_1\ \hat{h}_2\ \hat{h}_3\ \hat{h}_4]^\top$$

of $\mathbf{h} = h\sqrt{\frac{\mathrm{E}}{2}} [\mathrm{e}^{+\mathrm{j}\Phi_0}\ \mathrm{e}^{-\mathrm{j}\Phi_0}\ \mathrm{e}^{+\mathrm{j}\Phi_1}\ \mathrm{e}^{-\mathrm{j}\Phi_1}]^\top$ using standard least-square techniques. Thus, the coherent maximum likelihood (ML) symbol-by-symbol detector is given by

$$b_i^{\mathrm{ML}} = \arg\max_{b_i \in \{0,1\}} \exp\left\{ -\left\| \mathbf{r} - \hat{\mathbf{h}} \odot \mathbf{s}_{b_i} \right\|^2 \right\} \Leftrightarrow \tag{16}$$

$$\mathcal{Re}((\hat{h}_1)^* r_0^+ + (\hat{h}_2)^* r_0^-) \overset{\text{bit } 0}{\geq} \mathcal{Re}((\hat{h}_3)^* r_1^+ + (\hat{h}_4)^* r_1^-), \tag{17}$$

where $\mathcal{Re}(\cdot)$ stands for the real part. Equation (17) can be easily modified to offer ML coherent decoding; the latter was tested with RM and BCH channel-encoded sequences, both in simulation, as well as experimental setups [25].

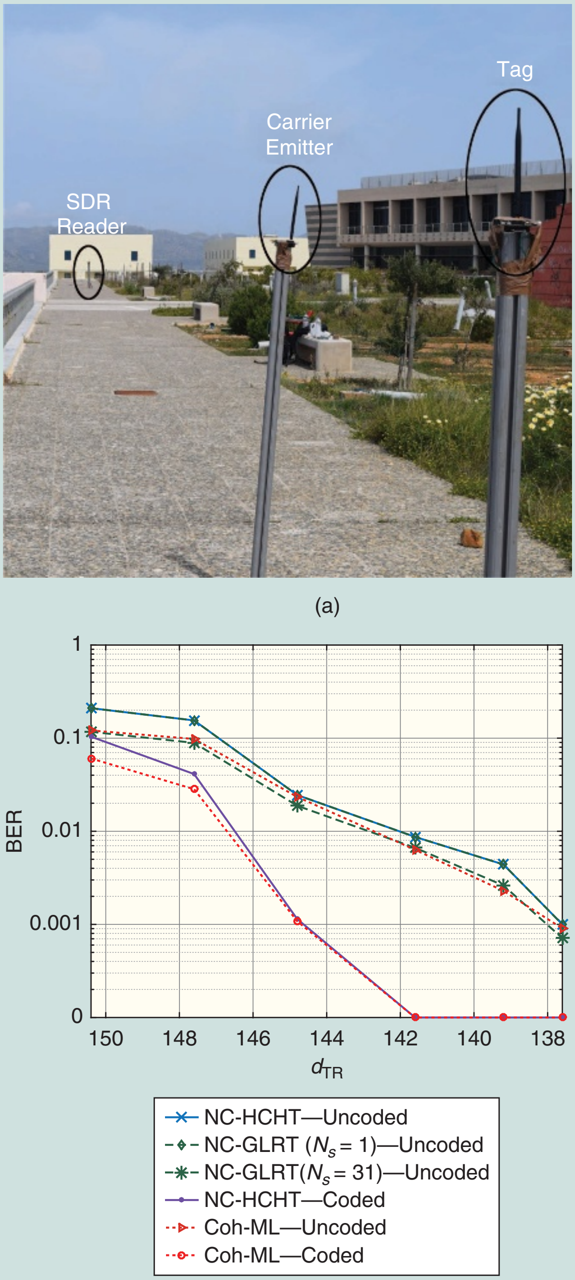
As previously mentioned in the section "Short-Packet/Sequence Detection–No Channel Coding," FM0 line coding, utilized in industrial (Gen2) RFID, can be seen as orthogonal signaling, despite the fact that industrial RFIDs utilize OOK and not FSK. Work in [26] exploited this interpretation, in conjunction with the 6-bit preambles already present in Gen2, estimated the channel, and performed optimal coherent detection with orthogonal signaling. Signal processing software for a GNU radio-based SDR receiver for Gen2/FM0 RFIDs was also open sourced.

Partially coherent detection

When all wireless channel-specific parameters are unknown but the receiver only has partial information regarding the tag-modulating phases Φ_0, Φ_1, the following partially coherent detector for backscatter FSK is possible [9]:

$$\left| r_0^+ + e^{2j\Phi_0} r_0^- \right| \overset{\text{bit 0}}{\geq} \left| r_1^+ + e^{2j\Phi_1} r_1^+ \right|, \qquad (18)$$

where $\mathbf{r} = [r_0^+ \; r_0^- \; r_1^+ \; r_1^-]^\top$ is defined as before and the receiver must know the tag-dependent, modulating phases Φ_0 and Φ_1; notice that this detector is different than the fully noncoherent square-law detector, described in the section "Symbol-by-Symbol Detection." Performance of the detector in (18) is given later in the section "Architectures and Network Applications."



(a)

(b)

FIGURE 6. (a) and (b) Intelligent signal processing with SDR, even with a noisy receiver, has been experimentally tested, with extended tag-to-reader (d_{TR}) distances. Parameters for the figure plot include P_C = 13 dBm, receiver NF $\in [7-12]$ dB, illuminator-to-tag distance d_{CT} = 8 m, T = 1 ms, $F_1 = 2F_0$ = 250 kHz, and 16 training (preamble) + 31 data-coded bits [20].

Comparison of coherent versus noncoherent short-packet detection

The major disadvantage of coherent communication is the utilization of preamble bits at the packet, a priori known at the reader, for channel estimation. In short-packet communication, e.g., with packet payload of only 32 channel-coded bits, a preamble of 8–16 bits is comparable to the payload, requiring comparable energy, deceasing the rate, and suggests inefficient communications. For batteryless tags, where every minuscule amount of power matters, such inefficiency is further amplified.

Work in [19] studied low-bit-rate backscatter FSK communication, comparing noncoherent HCHT versus coherent ML symbol-by-symbol detection, for small packets (on the order of 100 bits), under fixed energy per packet at both cases, i.e., taking into account the energy utilized for preamble bits (in the coherent case) in packet energy budget; no fading (AWGN), Rice and Rayleigh fading were studied. It was found that the BER performance gap between noncoherent and coherent was on the order of 1 dB or less, with decreasing value when going from Rayleigh to Rice to AWGN, i.e., from a more random to a more deterministic channel.

Experimental results in [20] with both symbol-by-symbol or sequence detection, using noncoherent or coherent techniques (including BCH and RM channel coding), as described in this work, corroborated such a (perhaps major) finding: noncoherent detection can be as good as coherent (Figure 6), boosting tag-to-reader communication distances, even with high noise-figure (NF) radios. For the case of noncoherent detection, synchronization was performed without any type of pilots/preambles, solely based on energy techniques.

A note on embedded receivers

In bistatic setups, highly sensitive, conventional (Marconi), embedded receivers can be utilized for backscatter radio reception. In that case, the tag must transmit the necessary protocol bits before (and after) payload that the embedded receiver is expecting. Given that radio sensitivity depends on communication bandwidth (with higher bandwidth resulting in lower sensitivity), NF, temperature of operation, and detection method and required minimum SNR, embedded receivers with small NF and small bandwidth can, in principle, detect signals with power below –100 dBm.

Examples of backscatter bistatic radio reception using embedded FSK radios include work in [27], where Bluetooth Low Energy (BLE)-embedded modules were utilized. Another recent example is offered in [28], where backscatter bistatic FSK was received by SI1064 or TI CC1101 embedded radios, with transmit power + 13 dBm at the illuminator and illuminator-to-tag, tag-to-embedded receiver distances at 3 m and 268 m, respectively, at packet error rate (PER) of 10.6% (Figure 7), or illuminator-to-tag, tag-to-embedded receiver distances at 3 m and 246 m, respectively, at ~1% PER. Subsequent work with LORA embedded receivers offered additional ranges with 20-dB additional illuminator transmission power and about 30-dB higher sensitivity (due to smaller bandwidth), compared to [28]. Note that 100 times smaller reception bandwidth results in 20-dB higher radio sensitivity.

Extensions to ambient environments

Theoretical studies

In [29], ambient backscatter communication is studied from an information-theoretic point of view. The setup includes an orthogonal frequency-division multiplexing (OFDM) ambient illuminator, the backscatter tag, and a legacy receiver for the illuminating signal; a dedicated receiver for the backscattered signal is considered as part of a second setup. Interestingly, it is shown that the additional paths created from the tag's backscattering may offer a performance gain for the legacy receiver.

In [30], differential modulation in conjunction with OOK is employed at the tag, while the ambient illuminator's complex baseband samples are considered to follow complex normal distribution; 8-PSK illumination is studied as well. Utilizing the signal model $y[n] = h_{\mathrm{CR}}\mathsf{c}[n] + h_{\mathrm{CT}}h_{\mathrm{TR}}\mathsf{s}B[n]\,\mathsf{c}[n] + w[n]$ [which follows (7)], where $\mathsf{c}[n]$ denotes the ambient carriers' complex samples s.t. $\mathsf{c}[n] \sim \mathcal{CN}(0, P_s)$, $B[n] \in \{0, 1\}$, the differentially encoded tag's signal, and $w[n] \sim \mathcal{CN}(0, N_w)$ additive noise, the following two hypotheses are formed:

$$y[n] = \begin{cases} \mathcal{CN}(0, \sigma_0^2), & B[n] = 0 \\ \mathcal{CN}(0, \sigma_1^2), & B[n] = 1, \end{cases} \tag{19}$$

where $\sigma_0^2 = |h_{\mathrm{CR}}|^2 P_s + N_w, \sigma_1^2 = |h_{\mathrm{CR}} + h_{\mathrm{CT}}h_{\mathrm{TR}}\mathsf{s}|^2 P_s + N_w$. Based on the two hypotheses, ML (based on the aforementioned signal model) and energy-based (based on summing $|y[n]|^2$ over the duration of a single bit and comparing with a threshold) detectors were derived. Both detectors required knowledge of ambient illuminator-and channel-related parameters σ_0^2, σ_1^2, acquired in a blind way with variance estimation. However, in the aforementioned detection method, a received sequence of the next/previous $M-1$ symbols is needed before detecting symbol M, with all channel-related parameters assumed unchanged for M bit periods. Using blind estimation, complex normal illumination and energy-based detection, BER of $\simeq 8 \cdot 10^{-3}$ was achieved at transmit (based on the ambient illuminator's power) SNR of 20 dB, complementing related work in [31]. Using a similar methodology, the authors in [32] omitted the differential encoding and employed a short training sequence to assist the blind estimation method, suggesting partially coherent detection. In [33], the repeating structure of an ambient OFDM carrier, due to the presence of cyclic prefix and the channel's effect, was exploited to derive an ML detector for a single antenna receiver; multiantenna receiver design was also studied. Modeling the ambient illuminator baseband signal as a complex Gaussian ignores the modulation format of the ambient signal; furthermore, performance of tag-backscattered signal detection on top of an ambient modulated carrier should take into account realistic channel conditions, transmission power, and link budgets.

Work in [34] considers a cognitive radio network (CRN) where the secondary system's transmitter (ST) is able to

1) utilize ambient backscatter under illumination from a primary transmitter (PT) toward a secondary receiver (SR)
2) harvest energy from PT transmissions
3) communicate with an SR using active radio powered by the harvested energy.

The authors, study both underlay (the primary channel is always busy) and overlay scenarios. Optimal (with respect to secondary rate) tradeoffs regarding time allocation between backscattering and energy harvesting are presented. For the underlay case, the rate optimization problem includes the constraint of maximum allowable ST transmission power to avoid interference toward the primary channel. All of the aforementioned are recent example efforts in this exciting, rapidly evolving field [35]. A contemporary survey can be found in [36].

Practical implementations

Implementation of an ambient backscatter communication system can be found in [18], where the authors exploited

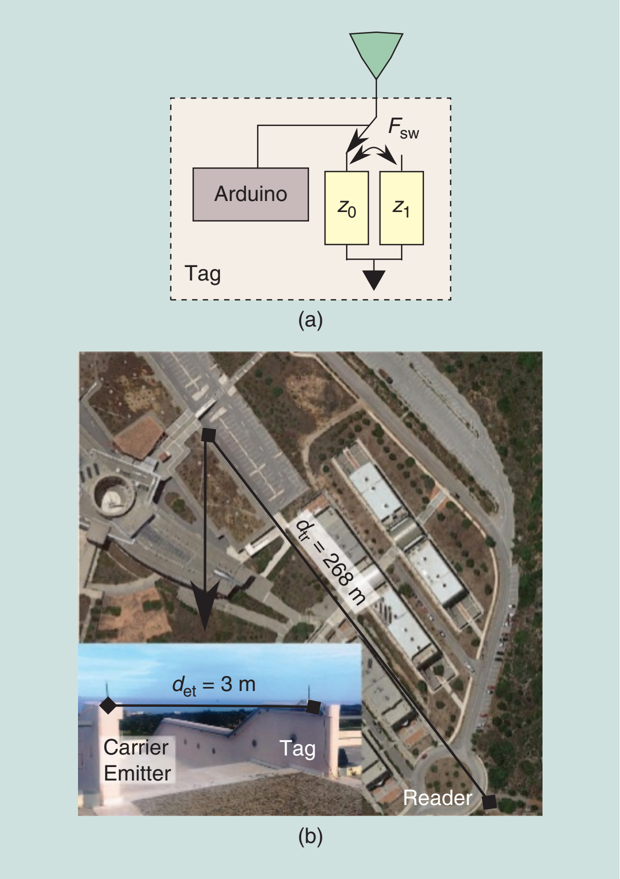


FIGURE 7. (a) A simple backscatter radio tag. (b) An experimental setup with an embedded radio receiver and illuminator-to-tag (d_{et}) and tag-to-embedded receiver (d_{tr}) distances at 3 m and 268 m, respectively [28].

illumination from ambient DTV signals and envelope detection/averaging to achieve tag-to-tag communication with range on the order of 60 cm. In [37], the authors exploited spread spectrum techniques, implemented in an analog, low-power fashion, to extend the range of tag-to-tag communication to (indicatively) 6 m for an impinged power of –15 dBm and bit rate of 3.3 bits/s, under illumination from DTV. In a similar manner, multiantenna analog design offered rates up to 1 Megabit/s with a communication range of 2 m, exploiting an impinged DTV power of –10 dBm.

Work in [38] exploited illumination from ambient FM radio signals, and the communication range (tag-to-FM receiver) was increased to approximately 18 m. Digital (audio 2-FSK, audio 4-FSK) as well analog (audio) communication was achieved. The tag was implemented using a function generator and a computer, while an integrated complementary metal–oxide–semiconductor (CMOS) design, implementing the functionality of the previous setup, was simulated. The same methodology was independently reported in [5], additionally providing a full prototype implementation [Figure 8(c)], consuming only 24 μW in continuous, nonduty-cycle operation and achieving a tag-to-receiver range of 26 m by exploiting selection diversity among various FM broadcasters; such selection diversity is easy to implement since the tag modulates directly at passband, and, thus, all FM broadcasting stations impinging on the tag antenna can be in principle exploited.

The latter two works previously mentioned demonstrate that an appropriate switching method, implemented at the tag, can result in minimum signal processing requirements at the reader side. Specifically, assume that the tag is illuminated by a FM-modulated signal, described as follows:

$$\mathsf{c}(t) = A_c \cos\left(2\pi F_c t + 2\pi k_s \int_0^t \phi(\tau) d\tau\right), \qquad (20)$$

where A_c is the carrier's amplitude, F_c is the carrier's center frequency, and $\phi(t)$ is the station's information (e.g., music). The tag RF switch is driven by an FM-modulated signal $\mathrm{x}_{\mathrm{sw,FM}}(t) = A_{\mathrm{sw}} \cos\left(2\pi F_{\mathrm{sw}} t + 2\pi k_{\mathrm{sw}} \int_0^t \mu(\tau) d\tau\right)$, where $\mu(t)$ is the tag information (e.g., from a sensor). As stated in the section "Universal System Model: Monostatic Versus Bistatic Versus Ambient," the backscattered signal (ignoring microwave-related parameters, noise, and fading terms for ease of explanation) takes the following form:

$$\begin{aligned} y_{\mathrm{bs}}(t) &= \mathrm{s}\mathsf{c}(t)\mathrm{x}_{\mathrm{sw,FM}}(t) \\ &= \frac{\gamma_s}{2}\cos(2\pi(F_c + F_{\mathrm{sw}})t + \Phi_s(t) + \Phi_{\mathrm{tag}}(t)) \\ &\quad + \frac{\gamma_s}{2}\cos(2\pi(F_c - F_{\mathrm{sw}})t + \Phi_s(t) - \Phi_{\mathrm{tag}}(t)), \qquad (21) \end{aligned}$$

where $\gamma_s = \mathsf{s}\, A_c\, A_{\mathrm{sw}}$, $\Phi_s(t) = 2\pi k_s \int_0^t \phi(\tau) d\tau$ and $\Phi_{\mathrm{tag}}(t) = 2\pi k_{\mathrm{sw}} \int_0^t \mu(\tau) d\tau$. Equation (21) demonstrates that if the tag is

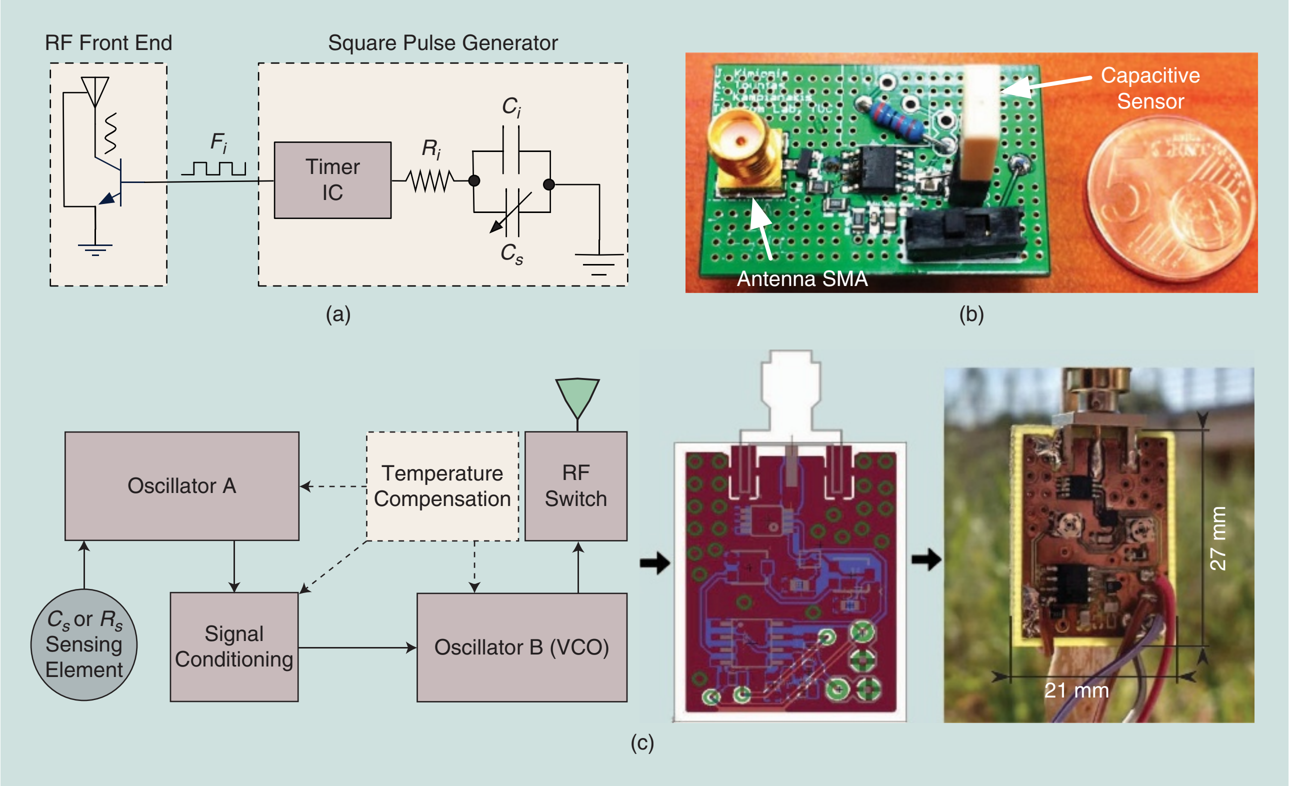


FIGURE 8. Ultralow-power backscatter radio tags: (a) and (b) environmental humidity [2]. (c) The resistive/capacitive sensor for ambient FM [5]. VCO: variable-control oscillator.

illuminated by an FM-modulated signal $\mathtt{c}(t)$ and the signal driving the RF switch is FM modulated as well (e.g., a square wave whose fundamental frequency is modulated according to the value of a sensor), then the backscattering operation results in two new FM-modulated signals, each centered at $F_s \pm F_{\mathrm{sw}}$. Thus, any receiver capable of performing FM demodulation can recover tag/sensors' signal $\mu(t)$ with, however, interference from illuminating stations' $\phi(t)$. Additionally, if $\mu(t)$ is limited in the audible frequency range, any conventional FM broadcast receiver, including modern smartphones, can recover $\mu(t)$. Using a method similar to [18], work in [39] also exploited ambient FM illumination, achieving ranges on the order of 5 m while consuming 1.78 mW for a bit rate of 1 kb/s in duty-cycled operation. Finally, Wi-Fi-based, ambient backscatter implementations can be found in [40]–[42].

Exploiting different illuminating signals involves different tradeoffs, depending on the ambient signal modulation, the technique used at the tag to remodulate information, and the receiver architecture. For example, exploiting TV signals in [18] required envelope detection, at the expense of limited communication ranges. In contrast, exploiting FM signals [5], [38] allowed for recovery of the backscattered information by any conventional FM receiver, while providing extended ranges and means for frequency-based multiuser communication, at the expense of more complicated (but widely available) FM signal demodulation.

Architectures and network applications

Monostatic versus bistatic/multistatic architectures

In certain applications, there is great need to maximize reliability and coverage. Thus, it is important to have a concrete network design principle, tailored to backscatter radio. More specifically, is it better to adopt a monostatic architecture, where illuminator and reader antenna are the same? Or is it better to adopt a bistatic architecture, where reader and illuminator are separated units, distant in space?

It turns out that, in terms of link budget, i.e., large-scale path loss, the inherent asymmetry of the bistatic architecture helps and the bistatic outperforms the monostatic architecture; assuming free-space loss (where the received power drops with the squared distance), fixed illuminator-to-reader distance $\mathsf{d}_{\max}$ and denote as x the illuminator-to-tag distance; it can be easily seen that the average received power at the reader is proportional to

$$y(x) = \left(\frac{1}{x}\right)^2 \left(\frac{1}{\mathsf{d}_{\max} - x}\right)^2,$$

which is minimized for $x = \mathsf{d}_{\max}/2$, i.e., when the tag is equidistant from the illuminator and reader antenna, which is the case in the monostatic architecture.

It also turns out that the bistatic architecture outperforms the monostatic (where, in the latter, a common antenna for transmit and receive is assumed) in terms of small-scale loss, i.e., fading-relevant metrics, such as diversity order. The following theorems state formally the elements highlighted previously, further assuming fading amplitude distributed according to Nakagami, with normalized (equal to one) average squared value [9]:

Theorem 5

Under Nakagami fading, the BER of monostatic architecture (illuminator and reader share the same antenna) with ML coherent detection of backscatter FSK is bounded as follows:

$$\mathbb{P}(e_{l,n}^{[\mathrm{m}]}) \leq \frac{1}{2}\left(\frac{\mathtt{M}_n + \mathtt{M}_n^2}{2\,\mathtt{SNR}_n^{[\mathrm{m}]}}\right)^{\frac{\mathtt{M}_n}{2}} \mathsf{U}\left(\frac{\mathtt{M}_n}{2}, \frac{1}{2}, \frac{\mathtt{M}_n + \mathtt{M}_n^2}{2\,\mathtt{SNR}_n^{[\mathrm{m}]}}\right), \qquad (22)$$

where $\mathtt{M}_n$ is the Nakagami parameter for link TR, $\mathsf{U}(\cdot,\cdot,\cdot)$ is given in [10, eq. (13.4.4)], and $\mathtt{SNR}_n^{[\mathrm{m}]}$ is the average received SNR for monostatic system. For Rayleigh fading ($\mathtt{M}_n = 1$), the diversity order is 1/2.

The BER bound (22) coincides with the performance of partially coherent envelope monostatic backscatter FSK detector of (18).

Theorem 6

Under dyadic Nakagami fading, the BER of a bistatic architecture (illuminator and reader are distinct units with different antennas, and respective links with tag are independent) with ML coherent detection of backscatter FSK is bounded as follows:

$$\mathbb{P}(e_{l,n}^{[\mathrm{b}]}) \leq \frac{1}{2}\left(\frac{2\mathtt{M}_{ln}\mathtt{M}_n}{\mathtt{SNR}_{l,n}^{[\mathrm{b}]}}\right)^{\mathtt{M}_n} \mathsf{U}\left(\mathtt{M}_n, 1 + \mathtt{M}_n - \mathtt{M}_{ln}, \frac{2\mathtt{M}_{ln}\mathtt{M}_n}{\mathtt{SNR}_{l,n}^{[\mathrm{b}]}}\right), \qquad (23)$$

where $\mathtt{M}_n$ and $\mathtt{M}_{ln}$ are the Nakagami parameters for links TR and CT, respectively, while $\mathtt{SNR}_{l,n}^{[\mathrm{b}]}$ is the average received SNR for bistatic system. Under dyadic Rayleigh fading ($\mathtt{M}_n = \mathtt{M}_{ln} = 1$), the diversity order is one.

The previous BER bound (23) coincides with the performance of the partially coherent envelope bistatic backscatter FSK detector of (18).

Theorems 5 and 6 show that the diversity order of the bistatic architecture (for Rayleigh fading) is twice that of the monostatic, due to the independence between illuminator-to-tag and tag-to-reader links, contrary to the monostatic case. Furthermore, Theorems 5 and 6 quantify BER for both noncoherent as well as coherent backscatter FSK; it can be shown that the bistatic achitecture prevails [9]. That finding also suggests that using more than one illuminator, i.e., extending bistatic to multistatic architectures, would be highly beneficial.

In fact, a proof-of-concept, digital, multistatic backscatter radio wireless sensor network (WSN) with a single receiver, four low-cost emitters, and multiple ambiently powered, low-bit-rate tags, perhaps the first of its kind, was experimentally demonstrated in [9]. The illuminators utilized only 13-dBm transmission power in a TDMA fashion, covering an outdoor area of $3{,}500\ \mathrm{m}^2$. Proof-of-concept, analog multistatic backscatter radio WSN with a single receiver and two low-cost emitters was presented in [2] for greenhouse environmental humidity sensing; more details are given next.

Backscatter wireless sensor networks

In the context of environmental sensing, backscatter networks have been developed for monitoring both environmental humidity and soil moisture. Work in [2] utilized analog backscatter principles based on (backscatter) FM modulation, with tags consuming 220−500 μW, while offering a root mean squared (RMS) error of 2% relative humidity, at ultralow cost (~3 Euro) per tag. The tags' implementation was based on capacitive sensing principles, where a change in a sensing capacitors' value, due to a variation in the sensed quantity (i.e., humidity), offered a change in the capacitance's dielectric constant, alternating the fundamental period of a timer [Figure 8(a) and (b)]; the latter simply controlled the frequency of switching at the tag antenna between two loads, thus shifting the backscattered signal frequency. Tags were deployed in a greenhouse and networked based on simple, frequency-division multiple access. Utilizing the same principles, work in [4] demonstrated soil moisture monitoring across a field with measurement RMS error of 1.9%, while consuming approximately ~100 − 200 μW, at a cost of ~5 Euro per tag; reduced power consumption was achieved by switching off circuit subcomponents when they were not used, while the sensing capacitor inserted in the ground was based on a custom design.

> **The major disadvantage of coherent communication is the utilization of preamble bits at the packet, a priori known at the reader, for channel estimation.**

How about using a plant as a battery and as a sensor? Work in [3] demonstrated the feasibility of implementing backscatter tags, able to measure and transmit (utilizing backscatter FM principles) the electric potential (EP) across two electrodes in the plants' stem, while being solely powered by the plant itself, using another pair of electrodes [Figure 9(a)]. A strong correlation was found between the EP signal, solar irradiation, and the time instants at which the plant was actually watered; thus, the backscattered EP signal indicated when the plant was actually watered (and not just the moisture level in the vicinity of the plant). The tag design consumed only 20 μW, while the plant could offer about 1 μW at noon time; therefore, duty-cycling was needed, allowing the tag to harvest sufficient energy from the plant before backscattering the (information-rich) EP signal.

The aforementioned implementations constitute realizations of backscatter links/networks, able to measure an environmental variable (e.g., humidity, soil moisture, EP of a plant) and transmit the value(s) toward an SDR under a dedicated illuminating carrier. As can be seen from the previous discussion, a computer running appropriate software (to decode the sensor's information from the backscattered signal) is needed along with a separate unit(s) providing the necessary illumination/carrier. Dedicated illumination is not required in ambient setups; work in [5] proposed a backscatter tag that is able to facilitate any capacitive or resistive sensor [Figure 8(c)] that backscatters its information toward any conventional FM radio receiver, including modern smartphones; capacitive soil moisture sensing for agriculture was demonstrated [Figure 9(b)], in more detail in the section "Extensions to Ambient Environments."

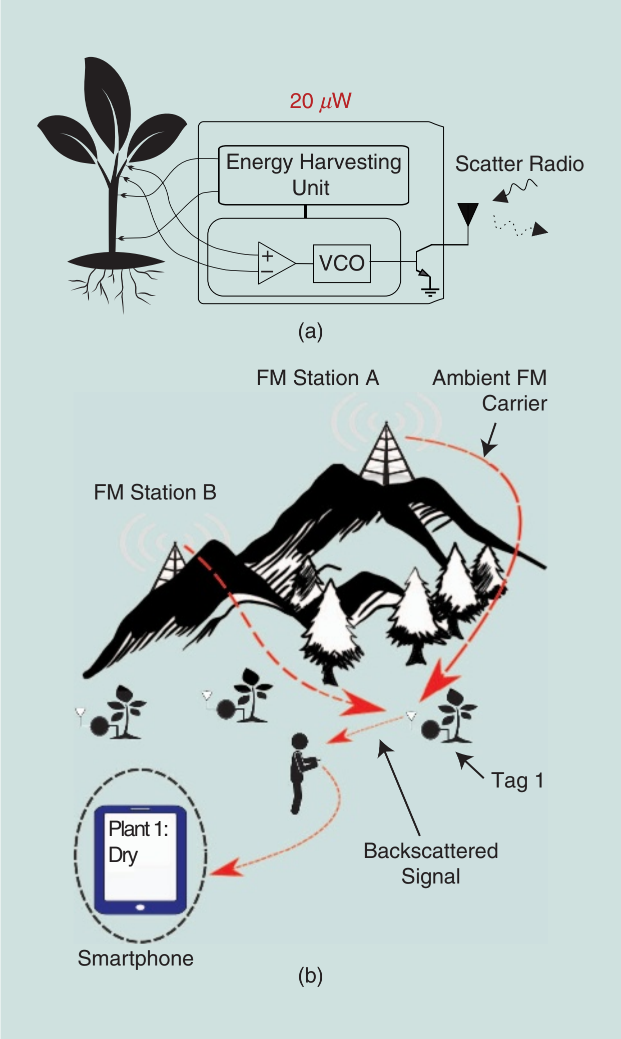


FIGURE 9. Backscatter radio-based IoT technologies could revolutionize environmental sensing and agriculture. (a) The backscattering of the EP of plants with μW consumption, powered by the plant itself [3]. (b) FM remodulation and backscattering from μW environmental sensors [5].

Discussion

Since switching between two antenna loads is the basic tag function for backscattering, a fundamental power consumption limit at each tag emerges: the power cost of a switch! Today's advanced CMOS technology operates at energies on the order of $10^4 - 10^5$ $k_B T_\theta$ per binary switching event using MOSFET switches and von-Neumann architectures, where k_B is the Boltzmann constant and T_θ is the temperature (in Kelvin). It is also noted that the fundamental limit of switching [Guardian Angels, FET Flagship Pilot, Final Report (public version), April 2012, based on information from R.W. Keyes, *IBM Journal of Research Development* vol. 32, 1988, pp. 24–28, data updated by T. Theis and R. Keyes, IBM Research, 2010] can be calculated from the Boltzmann probability, equal to $3\ln(2)k_BT_\theta \approx 10^{-21}$ J, at room temperature (300° K); assuming switching frequency at the tag between the two loads at 100 kHz, the aforementioned CMOS state-of-the-art energies of $10^4 - 10^5$ k_BT_θ correspond to $(0.5-5)\times 10^{-15}$W (fW) power consumption at the tag for room temperature.

The backscatter tag also requires power for the driving signal that controls switching and, of course, the rest of the circuitry needed for sensing (if sensing is also performed); examples of how backscatter radio and sensing can be performed jointly, with minimal additional hardware (e.g., adding a low-power timer), were previously given. Thus, the aforementioned numbers hint that further reduction of tag power consumption at the sub-μW regime is possible in the near future.

In the context of environmental sensing, backscatter networks have been developed for monitoring both environmental humidity and soil moisture.

In terms of signal processing, noncoherent sequence detection with relatively small complexity is challenging for ergodic setups, e.g., when the ambient illuminating signal changes for different backscatter tag bits, or nonergodic setups, when channel conditions (including ambient illuminator's signal) can be assumed constant during tag sequence backscattering; ideas from the work presented in this tutorial may assist. Initial results alongside this distinction, turning ambient modulated (but unknown) signals to an advantage, compared to unmodulated/CW illuminator, can be found in [43].

It is also important to realize that backscattering with simple switching, i.e., without amplifiers or another type of active signal conditioning at the tag, is, in principle, a communication technique at the power-limited regime; thus, typical housekeeping tasks for digital communications, such as CFO estimation, packet and symbol synchronization, dc offset removal, channel estimation, or more advanced tasks, such as successive interference cancellation, should not be idealized or overlooked at small SNRs. Furthermore, realistic assumptions on link budgets and noise at the receiver should be carefully justified.

Finally, it is perhaps important to note that backscatter radio was the first enabling technology for commercial RFID systems, realizing (even though with limited success) exploitation of a remote signal for both communication and power transfer; the same concept can be used to eliminate other parts in a typical receiver chain, e.g., remove the oscillator part in a receiver and exploit the carrier signal from a nearby transmitter (e.g., see the example work in [44]). Shared signaling and electronics open new avenues for realizable, ultralow-power IoT technology of the future. Other promising applications of backscatter radio (relevant to chipless RFID, motion detection, gesture recognition, and indoor localization but not covered in this work) underline the importance of backscatter radio in the years to come.

Acknowledgment

This research is implemented through the Operational Program "Human Resources Development, Education and Lifelong Learning" and is co-financed by the European Union (European Social Fund) and Greek national funds.

Authors

Aggelos Bletsas (aggelos@telecom.tuc.gr) received the diploma degree (with honors) in electrical and computer engineering from the Aristotle University of Thessaloniki, Greece, in 1998, and the S.M. and Ph.D. degrees in media arts and sciences from the Massachusetts Institute of Technology Media Lab, Cambridge in 2001 and 2005, respectively. He currently serves as an associate professor in the School of Electrical and Computer Engineering, Technical University of Crete, Greece. His research interests span the broad area of scalable wireless communications and sensors networking. He was a corecipient of the IEEE Communications Society 2008 Marconi Prize Paper Award in Wireless Communications, and various Best Student Paper Awards, e.g., at both the 2011 and 2017 IEEE International Conference on RFID-Technologies and Applications; the 2015 International Conference on Acoustics, Speech, and Signal Processing; and the 2018 International Conference on Modern Circuits and Systems Technologies.

Panos N. Alevizos (palevizos@isc.tuc.gr) received the diploma, M.Sc., and Ph.D. degrees in electronic and computer engineering from the Technical University of Crete, Greece, in 2012, 2014, and 2017, respectively. His research interests include communication theory and signal processing with an emphasis on backscatter radio and radio-frequency identification. He is a corecipient of the 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing Best Student Paper Award. He has been distinguished as an exemplary reviewer for 2015 by the editorial board of *IEEE Wireless Communications Letters* and for 2017 by the editorial board of *IEEE Transactions on Wireless Communications*.

Georgios Vougioukas (gevougioukas@isc.tuc.gr) received his five-year diploma degree in electrical and computer engineering from the Technical University of Crete, Greece, in 2016, and he is currently pursuing his Ph.D. degree in electrical and computer engineering there. His research interests include methods for ultralow-power wireless communication, signal processing for backscatter communication, energy harvesting, and analog and digital system design and implementation. He was a corecipient of the 2017 IEEE International Conference on RFID Technology and Applications Best Student Paper Award. He has been distinguished as an exemplary reviewer for 2017 by the editorial board of *IEEE Transactions on Wireless Communications*.

References

[1] G. Vannucci, A. Bletsas, and D. Leigh, "A software-defined radio system for backscatter sensor networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2170–2179, June 2008.

[2] E. Kampianakis, J. Kimionis, K. Tountas, C. Konstantopoulos, E. Koutroulis, and A. Bletsas, "Wireless environmental sensor networking with analog scatter radio & timer principles," *IEEE Sensors J.*, vol. 14, no. 10, pp. 3365–3376, Oct. 2014.

[3] C. Konstantopoulos, E. Koutroulis, N. Mitianoudis, and A. Bletsas, "Converting a plant to a battery and wireless sensor with scatter radio and ultra-low cost," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 2, pp. 388–398, Feb. 2016.

[4] S. N. Daskalakis, S. D. Assimonis, E. Kampianakis, and A. Bletsas, "Soil moisture scatter radio networking with low power," *IEEE Trans. Microw. Theory Techn.*, vol. 64, no. 7, pp. 2338–2346, July 2016.

[5] G. Vougioukas and A. Bletsas, "24μW 26m range batteryless backscatter sensors with FM remodulation and selection diversity," in *Proc. IEEE Radio Frequency Identification Technology and Applications*, Warsaw, Poland, Sept. 2017, pp. 237–242.

[6] S. J. Thomas and M. S. Reynolds, "A 96 Mbit/sec, 15.5 pJ/bit 16-QAM modulator for UHF backscatter communication," in *Proc. IEEE Radio Frequency Identification Conf.*, Orlando, FL, Apr. 2012, pp. 185–190.

[7] S. Thomas, E. Wheeler, J. Teizer, and M. Reynolds, "Quadrature amplitude modulated backscatter in passive and semipassive UHF RFID systems," *IEEE Trans. Microw. Theory Techn.*, vol. 60, no. 4, pp. 1175–1182, Apr. 2012.

[8] P. Hu, P. Zhang, and D. Ganesan, "Laissez-faire: Fully asymmetric backscatter communication," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 255–267, Oct. 2015.

[9] P. N. Alevizos, K. Tountas, and A. Bletsas, "Multistatic scatter radio sensor networks for extended coverage," *IEEE Trans. Wireless Commun.*, to be published.

[10] V. Iyer, V. Talla, B. Kellogg, S. Gollakota, and J. Smith, "Inter-technology backscatter: Towards internet connectivity for implanted devices," in *Proc. ACM Special Interest Group on Data Communications Conf.*, Florianopolis, Brazil, 2016, pp. 356–369.

[11] J. Kimionis, A. Bletsas, and J. N. Sahalos, "Increased range bistatic scatter radio," *IEEE Trans. Commun.*, vol. 62, no. 3, pp. 1091–1104, Mar. 2014.

[12] A. Bletsas, A. G. Dimitriou, and J. Sahalos, "Improving backscatter radio tag efficiency," *IEEE Trans. Microw. Theory Techn.*, vol. 58, no. 6, pp. 1502–1509, June 2010.

[13] J. Kimionis, A. Bletsas, and J. N. Sahalos, "Design and implementation of RFID systems with software defined radio," in *Proc. IEEE European Conf. Antennas and Propagation*, Prague, Czech Republic, Mar. 2012, pp. 3464–3468.

[14] J. Kimionis, A. Bletsas, and J. N. Sahalos, "Bistatic backscatter radio for tag read-range extension," in *Proc. IEEE Radio Frequency Identification Technology and Applications Conf.,* Nice, France, Nov. 2012, pp. 356–361.

[15] J. Kimionis, A. Bletsas, and J. N. Sahalos, "Bistatic backscatter radio for power-limited sensor networks," in *Proc. IEEE Global Communication Conf.*, Atlanta, GA, Dec. 2013, pp. 353–358.

[16] J. D. Griffin and G. D. Durgin, "Gains for RF tags using multiple antennas," *IEEE Trans. Antennas Propag.*, vol. 56, no. 2, pp. 563–570, Feb. 2008.

[17] J. D. Griffin and G. D. Durgin, "Complete link budgets for backscatter-radio and RFID systems," *IEEE Antennas Propag. Mag.*, vol. 51, no. 2, pp. 11–25, Apr. 2009.

[18] V. Liu, A. Parks, V. Talla, S. Gollakota, D. Wetherall, and J. R. Smith, "Ambient backscatter: Wireless communication out of thin air," in *Proc. ACM Special Interest Group on Data Communications Conf.*, Hong Kong, China, 2013, pp. 39–50.

[19] P. N. Alevizos and A. Bletsas, "Noncoherent composite hypothesis testing receivers for extended range bistatic scatter radio WSNs," in *Proc. IEEE Int. Conf. Communications*, London, June 2015, pp. 4448–4453.

[20] P. N. Alevizos, A. Bletsas, and G. N. Karystinos, "Noncoherent short packet detection and decoding for scatter radio sensor networking," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2128–2140, May 2017.

[21] P. N. Alevizos, Y. Fountzoulas, G. N. Karystinos, and A. Bletsas, "Noncoherent sequence detection of orthogonally modulated signals in flat fading with log-linear complexity," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, Brisbane, Australia, Apr. 2015, pp. 2974–2978.

[22] P. N. Alevizos, Y. Fountzoulas, G. N. Karystinos, and A. Bletsas, "Log-linear-complexity GLRT-optimal noncoherent sequence detection for orthogonal and RFID-oriented modulations," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1600–1612, Apr. 2016.

[23] N. Fasarakis-Hilliard, P. N. Alevizos, and A. Bletsas, "Coherent detection and channel coding for bistatic scatter radio sensor networking," *IEEE Trans. Commun.*, vol. 63, pp. 1798–1810, May 2015.

[24] P. N. Alevizos, N. Fasarakis-Hilliard, K. Tountas, N. Agadakos, N. Kargas, and A. Bletsas, "Channel coding for increased range bistatic backscatter radio: Experimental results," in *Proc. IEEE Radio Frequency Identification Technology and Applications Conf.*, Tampere, Finland, Sept. 2014, pp. 38–43.

[25] N. Fasarakis-Hilliard, P. N. Alevizos, and A. Bletsas, "Coherent detection and channel coding for bistatic scatter radio sensor networking," in *Proc. IEEE Int. Conf. Communications*, June 2015, pp. 4895–4900.

[26] N. Kargas, F. Mavromatis, and A. Bletsas, "Fully-coherent reader with commodity SDR for Gen2 FM0 and computational RFID," *IEEE Wireless Commun. Lett.*, vol. 4, no. 6, pp. 617–620, Dec. 2015.

[27] J. F. Ensworth and M. S. Reynolds, "Every smart phone is a backscatter reader: Modulated backscatter compatibility with Bluetooth 4.0 Low Energy (BLE) devices," in *Proc. IEEE Radio Frequency Identification Conf.*, San Diego, CA, Apr. 2015, pp. 78–85.

[28] G. Vougioukas, S. N. Daskalakis, and A. Bletsas, "Could battery-less scatter radio tags achieve 270-meter range?," in *Proc. IEEE Wireless Power Transfer Conf.*, Aveiro, Portugal, May 2016, pp. 1–3.

[29] D. Darsena, G. Gelli, and F. Verde, "Modeling and performance analysis of wireless networks with ambient backscatter devices," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1797–1814, Apr. 2017.

[30] J. Qian, F. Gao, G. Wang, S. Jin, and H. Zhu, "Noncoherent detections for ambient backscatter system," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1412–1422, Mar. 2017.

[31] G. Wang, F. Gao, R. Fan, and C. Tellambura, "Ambient backscatter communication systems: Detection and performance analysis," *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4836–4846, Nov. 2016.

[32] J. Qian, F. Gao, G. Wang, S. Jin, and H. Zhu, "Semi-coherent detection and performance analysis for ambient backscatter system," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5266–5279, Dec. 2017.

[33] G. Yang, Y. Liang, R. Zhang, and Y. Pei. (2017). Modulation in the air: Backscatter communication over ambient OFDM carrier. [Online]. Available: http://arxiv.org/abs/1704.02245

[34] D. T. Hoang, D. Niyato, P. Wang, D. I. Kim, and Z. Han, "Ambient backscatter: A new approach to improve network performance for RF-powered cognitive radio networks," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3659–3674, Sept. 2017.

[35] K. Han and K. Huang, "Wirelessly powered backscatter communication networks: Modeling, coverage and capacity," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2548–2561, Apr. 2017.

[36] N. V. Huynh, D. T. Hoang, X. Lu, D. Niyato, P. Wang, and D. I. Kim, "Ambient backscatter communications: A contemporary survey," arXiv Preprint, arXiv: 1712.04804, Dec. 2017.

[37] A. N. Parks, A. Liu, S. Gollakota, and J. R. Smith, "Turbocharging ambient backscatter communication," in *Proc. ACM Special Interest Group Data Communications Conf.*, Chicago, IL, 2014, pp. 619–630.

[38] A. Wang, V. Iyer, V. Talla, J. R. Smith, and S. Gollakota, "FM backscatter: Enabling connected cities and smart fabrics," in *Proc. USENIX Symp. Networked Systems Design and Implementation Conf.*, Boston, MA, Mar. 2017, pp. 243–258.

[39] S. N. Daskalakis, J. Kimionis, A. Collado, G. Goussetis, M. M. Tentzeris, and A. Georgiadis, "Ambient backscatterers using FM broadcasting for low cost and low power wireless applications," *IEEE Trans. Microw. Theory Techn.*, vol. PP, no. 99, pp. 1–12, 2017.

[40] B. Kellogg, A. Parks, S. Gollakota, J. R. Smith, and D. Wetherall, "Wi-Fi backscatter: Internet connectivity for rf-powered devices," in *Proc. ACM Special Interest Group Data Communications Conf.*, Chicago, IL, 2014, pp. 607–618.

[41] D. Bharadia, K. R. Joshi, M. Kotaru, and S. Katti, "Backfi: High throughput WiFi backscatter," in *Proc. ACM Special Interest Group Data Communications Conf.*, London, 2015, pp. 283–296.

[42] P. Zhang, D. Bharadia, K. Joshi, and S. Katti, "Hitchhike: Practical backscatter using commodity WiFi," in *Proc. ACM Conf. Embedded Networked Sensing Systems*, 2016, pp. 259–271.

[43] G. Vougioukas and A. Bletsas, "Switching frequency techniques for universal ambient backscatter networking," submitted for publication.

[44] A. Varshney, O. Harms, C. P. Penichet, C. Rohner, F. Hermans, and T. Voigt, "Lorea: A backscatter architecture that achieves a long communication range," in *Proc. ACM Conf. Embedded Networked Sensing Systems*, Delft, The Netherlands, Nov. 2017, pp. 50:1–50:2.

SP

Liang Xiao, Xiaoyue Wan, Xiaozhen Lu,
Yanyong Zhang, and Di Wu

IoT Security Techniques Based on Machine Learning

How do IoT devices use AI to enhance security?

INTERNET OF THINGS—ISTOCKPHOTO.COM/IAREMENKO
CIRCUITS—IMAGE LICENSED BY INGRAM PUBLISHING

The Internet of things (IoT), which integrates a variety of devices into networks to provide advanced and intelligent services, has to protect user privacy and address attacks such as spoofing attacks, denial of service (DoS) attacks, jamming, and eavesdropping. We investigate the attack model for IoT systems and review the IoT security solutions based on machine-learning (ML) techniques including supervised learning, unsupervised learning, and reinforcement learning (RL). ML-based IoT authentication, access control, secure offloading, and malware detection schemes to protect data privacy are the focus of this article. We also discuss the challenges that need to be addressed to implement these ML-based security schemes in practical IoT systems.

Introduction

The IoT facilitates integration between the physical world and computer communication networks, and applications (apps) such as infrastructure management and environmental monitoring make privacy and security techniques critical for future IoT systems [1]–[3]. Consisting of radio-frequency identifications (RFIDs), wireless sensor networks (WSNs), and cloud computing [4], IoT systems have to protect data privacy and address security issues such as spoofing attacks, intrusions, DoS attacks, distributed DoS (DDoS) attacks, jamming, eavesdropping, and malware [5], [6]. For instance, wearable devices that collect and send the user health data to a connected smartphone have to avoid privacy information leakage.

It's generally prohibitive for IoT devices with restricted computation, memory, radio bandwidth, and battery resources to execute computational-intensive and latency-sensitive security tasks, especially under heavy data streams [7]. However, most existing security solutions generate a heavy computation and communication load for IoT devices, and outdoor IoT devices such as cheap sensors with lightweight security protections are usually more vulnerable to attacks than computer systems. As shown in Figure 1, we investigate IoT authentication, access control, secure offloading, and malware detection.

Digital Object Identifier 10.1109/MSP.2018.2825478
Date of publication: 28 August 2018

1053-5888/18©2018IEEE

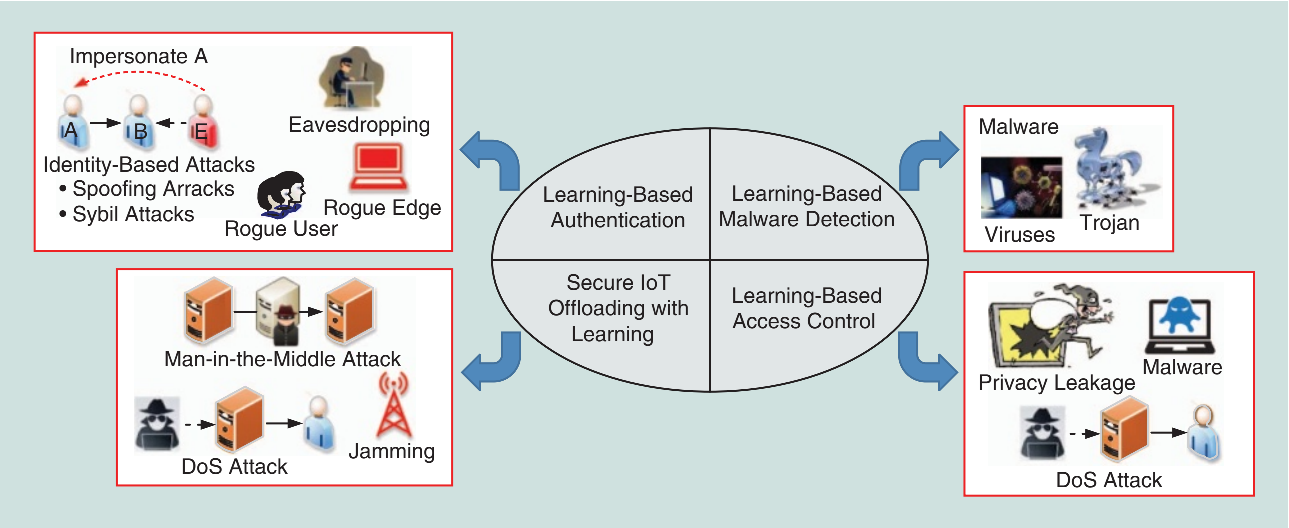


FIGURE 1. An illustration of the threat model in the IoT.

- Authentication helps IoT devices distinguish the source nodes and address identity-based attacks such as spoofing and Sybil attacks [8].
- Access control prevents unauthorized users from accessing the IoT resources [9].
- Secure offloading techniques enable IoT devices to use the computation and storage resources of the servers and edge devices for computational-intensive and latency-sensitive tasks [10].
- Malware detection protects IoT devices from privacy leakage, power depletion, and network performance degradation against malware such as viruses, worms, and Trojans [11].

With the development of ML and smart attacks, IoT devices have to choose a defensive policy and determine the key parameters in the security protocols for the tradeoff in the heterogenous and dynamic networks. This task is challenging as an IoT device with restricted resources usually has difficulty accurately estimating the current network and attack state in time. For example, the authentication performance of the scheme in [8] is sensitive to the test threshold in the hypothesis test, which depends on both the radio propagation model and the spoofing model. Such information is unavailable for most outdoor sensors, leading to a high false alarm or misdetection rate in the spoofing detection.

ML techniques including supervised learning, unsupervised learning, and RL have been widely applied to improve network security as summarized in Table 1, such as authentication, access control, antijamming offloading, and malware detection [8]–[22].

- Supervised learning techniques such as support vector machines (SVMs), naive Bayes, K-nearest neighbor (K-NN), neural networks (NNs), deep NNs (DNNs), and random forest can be used to label the network traffic or app traces of IoT devices to build the classification or regression model [9]. For example, IoT devices can use SVMs to detect network intrusion [9] and spoofing attacks [12], apply K-NNs in network intrusion [13] and malware [14] detection, and utilize NNs to detect network intrusion [15] and DoS attacks [16]. Naive Bayes can be applied by IoT devices in intrusion detection [9], and random forest classifier can be used to detect malware[14]. IoT devices with sufficient computation and memory resources can utilize DNNs to detect spoofing attacks [23].
- Unsupervised learning does not require labeled data in the supervised learning and investigates the similarity between the unlabeled data to cluster them into different groups [9]. For example, IoT devices can use multivariate correlation analysis to detect DoS attacks [17] and apply the infinite Gaussian mixture model (IGMM) in the physical (PHY)-layer authentication with privacy protection [18].
- RL techniques such as Q-learning, Dyna-Q, postdecision state (PDS) [24], and deep Q-network (DQN) [25] enable an IoT device to choose security protocols as well as key parameters against various attacks via trial and error [8]. For example, Q-learning as a model-free RL technique has been used to improve the performance of authentication [8], antijamming offloading [10], [19], [20], and malware detection [11], [21]. IoT devices can apply Dyna-Q in authentication and malware detection [11], use PDS to detect malware [11], and DQN in antijamming transmissions [22].

IoT attack model

Consisting of things, services, and networks, IoT systems are vulnerable to network, physical, and software attacks as well as privacy leakage. As shown in Figure 1, we focus on the IoT security threats as follows:

- *DoS attackers*: The attackers flood the target server with superfluous requests to prevent IoT devices from obtaining services [4]. One of the most dangerous types of a DoS attack is when DDoS attackers use thousands of Internet protocol addresses to request IoT services, making it difficult for the server to distinguish the legitimate IoT devices from attackers. Distributed IoT devices with lightweight security protocols are especially vulnerable to DDoS attacks [5].

- *Jamming*: Attackers send fake signals to interrupt the ongoing radio transmissions of IoT devices and further deplete the bandwidth, energy, central processing units (CPUs), and memory resources of IoT devices or sensors during their failed communication attempts [22].
- *Spoofing*: A spoofing node impersonates a legal IoT device with its identity such as the medium access control (MAC) address and RFID tag to gain illegal access to the IoT system and can further launch attacks such as DoS and man-in-the-middle attacks [8].
- *Man-in-the-middle attack*: A man-in-the-middle attacker sends jamming and spoofing signals with the goal of secretly monitoring, eavesdropping, and altering the private communication between IoT devices [4].
- *Software attacks*: Mobile malware such as Trojans, worms, and viruses can result in privacy leakage, economic loss, power depletion, and network performance degradation of IoT systems [11].
- *Privacy leakage*: IoT systems have to protect user privacy during data caching and exchange. Some caching owners are curious about the data content stored on their devices and analyze and sell such IoT privacy information. Wearable devices that collect user's personal information such as location and health information have witnessed an increased risk of personal privacy leakage [26].

Learning-based authentication

Traditional authentication schemes are not always applicable to IoT devices with limited computation, battery, and memory resources to detect identity-based attacks such as spoofing and Sybil attacks. PHY-layer authentication techniques that exploit the spatial decorrelation of the PHY-layer features of radio channels and transmitters such as the received signal strength indicators (RSSIs), received signal strength (RSS), channel impulse responses (CIRs) of the radio channels, channel state information (CSI), and the MAC address can provide lightweight security protection for IoT devices without leaking user privacy information [8].

PHY-layer authentication methods such as [8] build hypothesis tests to compare the PHY-layer feature of the message under test with the record of the claimed transmitter. Their authentication accuracy depends on the test threshold in the hypothesis test. However, it is challenging for an IoT device to choose an appropriate test threshold of the authentication due to the radio environment and the unknown spoofing model. The IoT device estimates the false alarm and misdetection rate of the spoofing detection at the last time slot, and the state of the learning consists of the false alarm and misdetection rate. The future state observed by the IoT device is independent of the previous states and actions if the current state and test threshold are known. Therefore, the test threshold selection in the IoT authentication in the repeated game against spoofing attacks can be viewed as a Markov decision process (MDP) with finite states.

The Q-learning-based authentication as proposed in [8] depends on the RSSI of the signals under test and enables an IoT device to achieve the optimal test threshold and improve the utility and the authentication accuracy. For example, the Q-learning-based authentication reduces the average authentication error rate by 64.3%, to less than 5%, and increases the utility by 14.7% compared with the PHY-authentication with a fixed threshold in an experiment performed in a $12 \times 9.5 \times 3$ m^3 lab with 12 transmitters [8].

Supervised learning techniques such as distributed Frank-Wolfe (dFW) and incremental aggregated gradient (IAG) can also be applied in IoT systems to improve spoofing resistance.

Table 1. ML-based IoT security methods.

Attacks	Security Techniques	ML Techniques	Performance
DoS	Secure IoT offloading	NN [16]	Detection accuracy
	Access control	Multivariate correlation analysis [17]	Root mean error
		Q-learning [21]	
Jamming	Secure IoT offloading	Q-learning [19], [20]	Energy consumption SINR
		DQN [22]	
Spoofing	Authentication	Q-learning [8]	Average error rate
		Dyna-Q [8]	Detection accuracy
		SVM [12]	Classification accuracy
		DNN [23]	False alarm rate
		dFW [27]	Missdetection rate
		Incremental aggregated gradient [27]	
Intrusion	Access control	SVM [9]	Classification accuracy
		Naive Bayes [9]	False alarm rate
		K-NNs [13]	Detection rate
		NN [15]	Root mean error
Malware	Malware detection Access control	Q/Dyna-Q/PDS [11]	Classification accuracy
		Random forest [14]	False positive rate
		K-NNs [14]	True positive rate
			Detection accuracy
			Detection latency
Eavesdropping	Authentication	Q-learning [10]	Proximity passing rate
		Nonparametric Bayesian [18]	Secrecy data rate

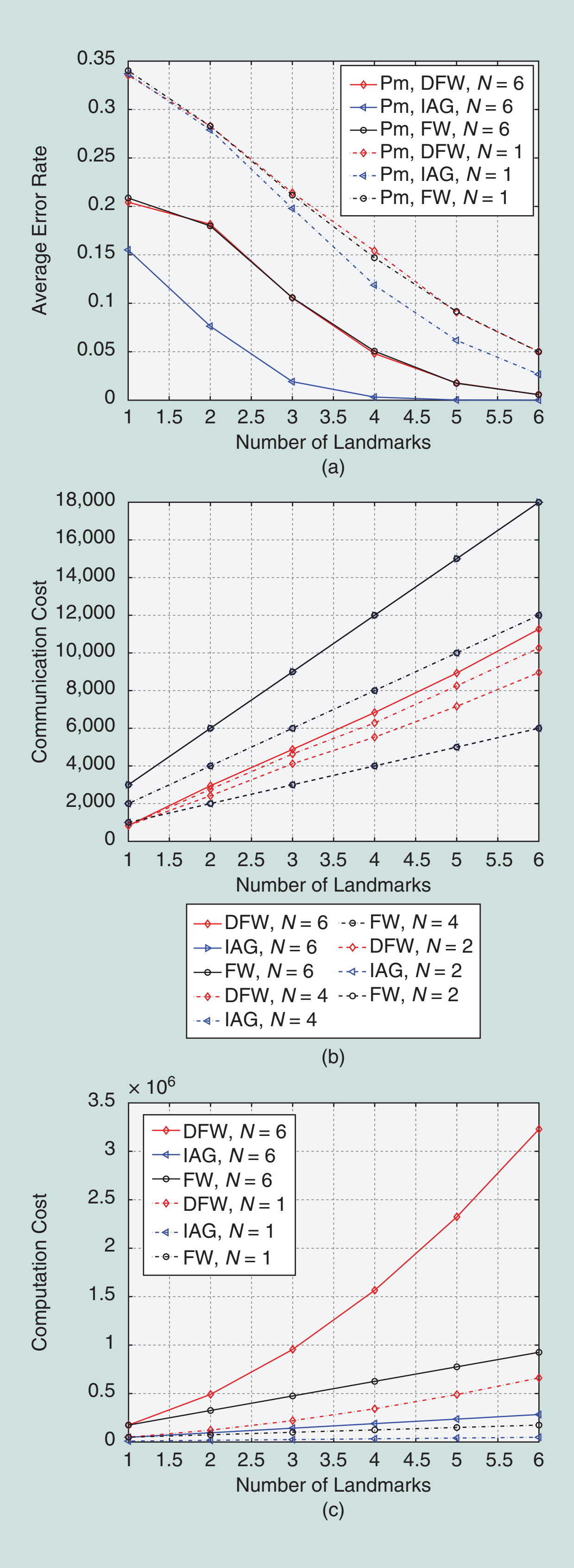


FIGURE 2. The performance of a PHY-layer authentication system with a different number of antennas at each landmark [27]: (a) average error rate, (b) communication cost, and (c) computation cost.

The authentication scheme in [27] exploits the RSSIs received by multiple landmarks and uses logistic regression to avoid being restricted to a known radio channel model. By applying the dFW and IAG algorithms to estimate the parameters of the logistic regression model, this authentication scheme saves communication overhead and improves spoofing detection accuracy. As shown in Figure 2, the average error rates of the dFW-based authentication and the IAG-based scheme are 6% and less than 10^{-4}, respectively, in the simulation with six landmarks, each equipped with six antennas. The dFW-based authentication reduces the communication overhead by 37.4%, while the IAG reduces the computation overhead by 71.3% compared with the Frank-Wolfe-based scheme in this case [27].

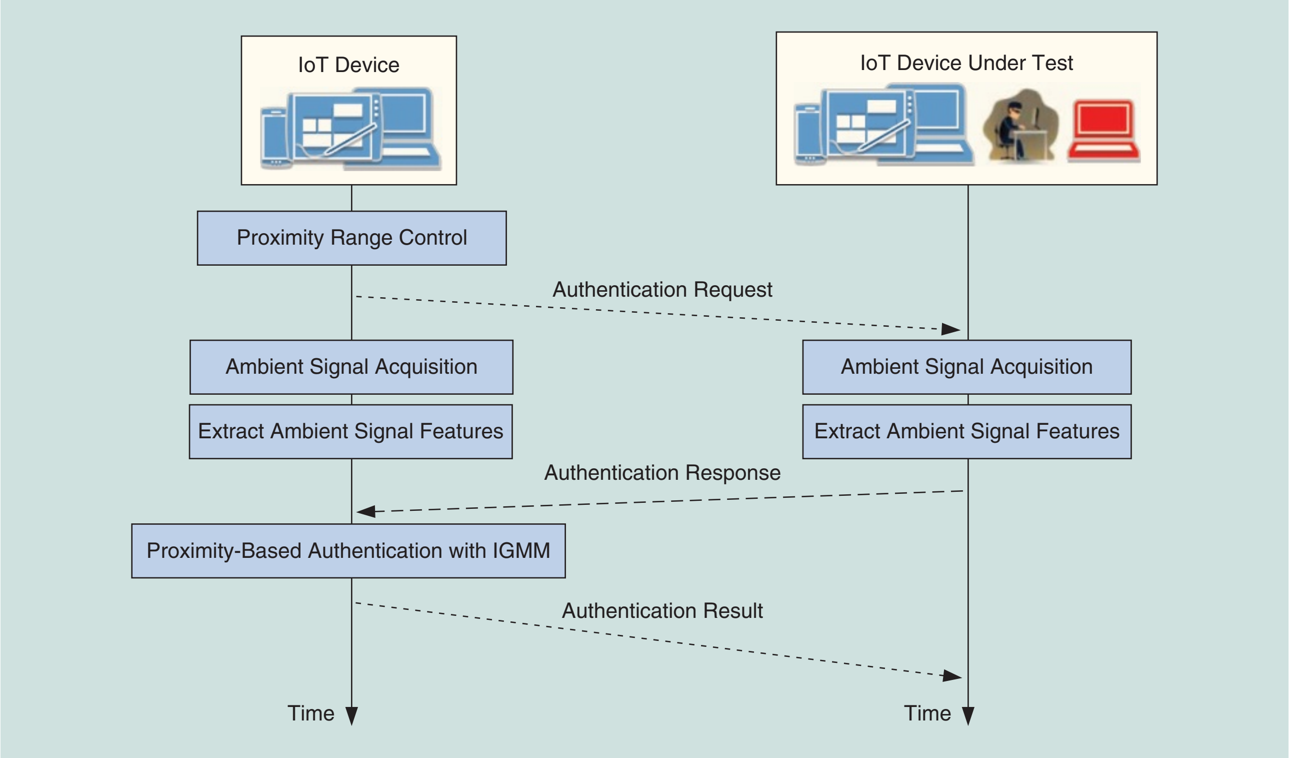
Unsupervised learning techniques such as IGMM can be applied in proximity-based authentication to authenticate the IoT devices in the proximity without leaking the localization information of the devices. For instance, the authentication scheme as proposed in [18] uses IGMM, a nonparametric Bayesian method to avoid the "overfitting" problem and, thus, adjust the model complexity, to evaluate the RSSIs and the packet arrival time intervals of the ambient radio signals to detect spoofers outside the proximity range. This scheme reduces the detection error rate by 20% to 5%, compared with the Euclidean distance-based authentication [18] in the spoofing detection experiments in an indoor environment.

As shown in Figure 3, this scheme requests the IoT device under test to send the ambient signals' features such as the RSSIs, MAC addresses, and packet arrival time interval of the ambient signals received during a specific time duration. The IoT device extracts and sends the ambient signals' features to the legal receiver. Upon receiving such authentication messages, the receiver applies IGMM to compare the reported signal features with those of the ambient signals observed in the proximity-based test. The receiver provides the IoT device passing the authentication with access to the IoT resources.

Finally, deep-learning techniques such as DNNs can be applied for IoT devices with sufficient computation and memory resources to further improve the authentication accuracy. The DNN-based user authentication as presented in [23] extracts the CSI features of the Wi-Fi signals and applies DNNs to detect spoofing attackers. The spoofing detection accuracy of this scheme is about 95%, and the user identification accuracy is 92.34% [23].

Learning-based access control

It is challenging to design access control for IoT systems in heterogeneous networks with multiple types of nodes and multisource data [9]. ML techniques such as SVMs, K-NNs, and NNs have been used for intrusion detection [15]. For instance, the DoS attack detection as proposed in [17] uses multivariate correlation analysis to extract the geometrical correlations between network traffic features. This scheme increases the detection accuracy by 3.05% to 95.2% compared with the triangle-area-based nearest-neighbors approach using the KDD Cup 99 data set [17].



IoT Device
IoT Device Under Test
Proximity Range Control
Authentication Request
Ambient Signal Acquisition
Ambient Signal Acquisition
Extract Ambient Signal Features
Extract Ambient Signal Features
Authentication Response
Proximity-Based Authentication with IGMM
Authentication Result
Time
Time

FIGURE 3. An illustration of ML-based authentication in IoT systems.

IoT devices such as outdoor sensors usually have strict resource and computation constraints, yielding challenges for anomaly intrusion detection techniques and thus degrading the intrusion detection performance for IoT systems. ML techniques help build lightweight access control protocols to save energy and extend the lifetime of IoT systems. For example, the outlier detection scheme as developed in [13] applies K-NNs to address the problem of unsupervised outlier detection in WSNs and offers flexibility to define outliers with reduced energy consumption. This scheme can save the maximum energy by 61.4% compared with the centralized scheme with similar average energy consumption [13].

The multilayer perceptron (MLP)-based access control as presented in [16] utilizes the NN with two neurons in the hidden layer to train the connection weights of the MLP and compute the suspicion factor that indicates whether an IoT device is the victim of DoS attacks. This scheme utilizes backpropagation (BP) that applies the forward computation and error BP and particle swarm optimization (PSO) as an evolutionary computation technique that utilizes particles with adjustable velocities to update the connection weights of the MLP. The IoT device under test shuts down the MAC- and PHY-layer functions to save energy and extend the network life if the output of the MLP exceeds a threshold.

Supervised learning techniques such as SVMs are used to detect multiple types of attacks for Internet traffic [28] and the smart grid [12]. For instance, a lightweight attack-detection mechanism as proposed in [28] uses an SVM-based hierarchical structure to detect traffic flooding attacks. In the attack experiment, the data set collector system gathered Simple Network Management Protocol (SNMP) management information base data from the victim system using SNMP query messages. Experiment results show that this scheme can achieve an attack detection rate over 99.40% and classification accuracy over 99.53% [28].

Secure IoT offloading with learning

IoT offloading has to address the attacks launched from the PHY- or MAC- layer attacks, such as jamming, rogue edge devices, rogue IoT devices, eavesdropping, man-in-the-middle attacks, and smart attacks [29]. As the future state observed by an IoT device is independent of the previous states and actions for a given state and offloading strategy in the current time slot, the mobile offloading strategy chosen by the IoT device in the repeated game with jammers and interference sources can be viewed as an MDP with finite states [10]. RL techniques can be used to optimize the offloading policy in dynamic radio environments.

Q-learning, as a model-free RL technique, is convenient to implement with low computation complexity. For example, IoT devices can utilize the Q-learning-based offloading as proposed in [10] to choose their offloading data rates against jamming and spoofing attacks. As illustrated in Figure 4, the IoT device observes the task importance, the received jamming power, the radio channel bandwidth, and the channel gain to formulate its current state, which is the basis to choose the offloading policy according to the Q-function. The Q-function is the expected discounted long-term reward for each action-state pair and

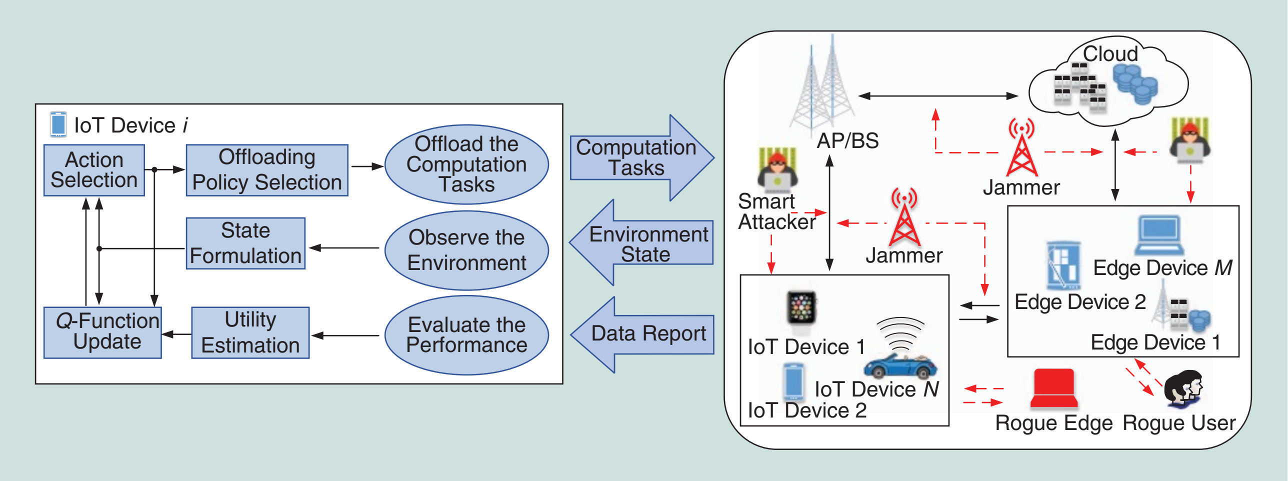


FIGURE 4. An illustration of ML-based offloading. AP: access point; BS: base station.

represents the knowledge obtained from the previous antijamming offloading. The Q-values are updated via the iterative Bellman equation in each time slot according to the current offloading policy, the network state, and the utility received by the IoT device against jamming.

The IoT device evaluates the signal-to-interference-plus-noise ratio (SINR) of the received signals, secrecy capacity, offloading latency, and energy consumption of the offloading process and estimates the utility in this time slot. The IoT device applies the ϵ-greedy algorithm in the offloading policy selection, in which the offloading policy with the max Q-value is selected with a high probability and the other policies are chosen with a small probability. Therefore, the IoT device makes a tradeoff between the exploration (i.e., to avoid being trapped in the local optimal strategy) and the exploitation (i.e., to improve the long-term reward). This scheme reduces the spoofing rate by 50% and decreases the jamming rate by 8% compared with a benchmark strategy as presented in [10].

According to the Q-learning-based antijamming transmission as proposed in [19], an IoT device can apply Q-learning to choose the radio channel to access the cloud or edge device without being aware of the jamming and interference model in IoT systems. As shown in Figure 4, the IoT device observes the center frequency and radio bandwidth of each channel to formulate the state and chooses the optimal offloading channel based on the current state and Q-function. Upon receiving the computation report, the IoT device evaluates the utility and updates the Q values. Simulation results in [19] show that this scheme increases the average cumulative reward by 53.8% compared with the benchmark random channel selection strategy.

Q-learning also helps IoT devices achieve the optimal subband from the radio spectrum band to resist jamming and interference from other radio devices. As shown in Figure 4, the IoT device observes the spectrum occupancy to formulate the state and selects the spectrum band accordingly. In an experiment against a sweeping jammer and in the presence of two wideband autonomous cognitive radios with ten subbands, this scheme increases the jamming cost by 44.3% compared with the benchmark subband selection strategy in [20].

The DQN-based antijamming transmission as developed in [22] accelerates the learning speed for IoT devices with sufficient computation and memory resources to choose the radio frequency channel. This scheme applies the convolutional NN (CNN) to compress the state space for large-scale networks with a large number of IoT devices and jamming policies in a dynamic IoT system and thus increase the SINR of the received signals. More specifically, the CNN consists of two convolutional layers and two fully connected layers. The weights of the CNN are updated based on the stochastic gradient descent algorithm according to the previous experience in the memory pool. The output of the CNN is used for estimating the values of the Q-function for each antijamming transmission policy. This scheme increases the SINR of the received signals by 8.3% and saves 66.7% of the learning time compared with the Q-learning scheme in the offloading against jamming attacks [22].

Learning-based IoT malware detection

IoT devices can apply supervised learning techniques to evaluate the runtime behaviors of the apps in malware detection. In the malware detection scheme as developed in [14], an IoT device uses K-NNs and random forest classifiers to build the malware-detection model. As illustrated in Figure 5, the IoT device filters the TCP packets and selects the features among various network features including the frame number and length, labels them, and stores these features in the database. The K-NN-based malware detection assigns the network traffic to the class with the largest number of objects among its K-NNs. The random forest classifier builds the decision trees with the labeled network traffic to distinguish malware. According to the experiments in [14], the true positive rates of the K-NN-based malware detection and random forest-based scheme with the MalGenome data set are 99.7% and 99.9%, respectively.

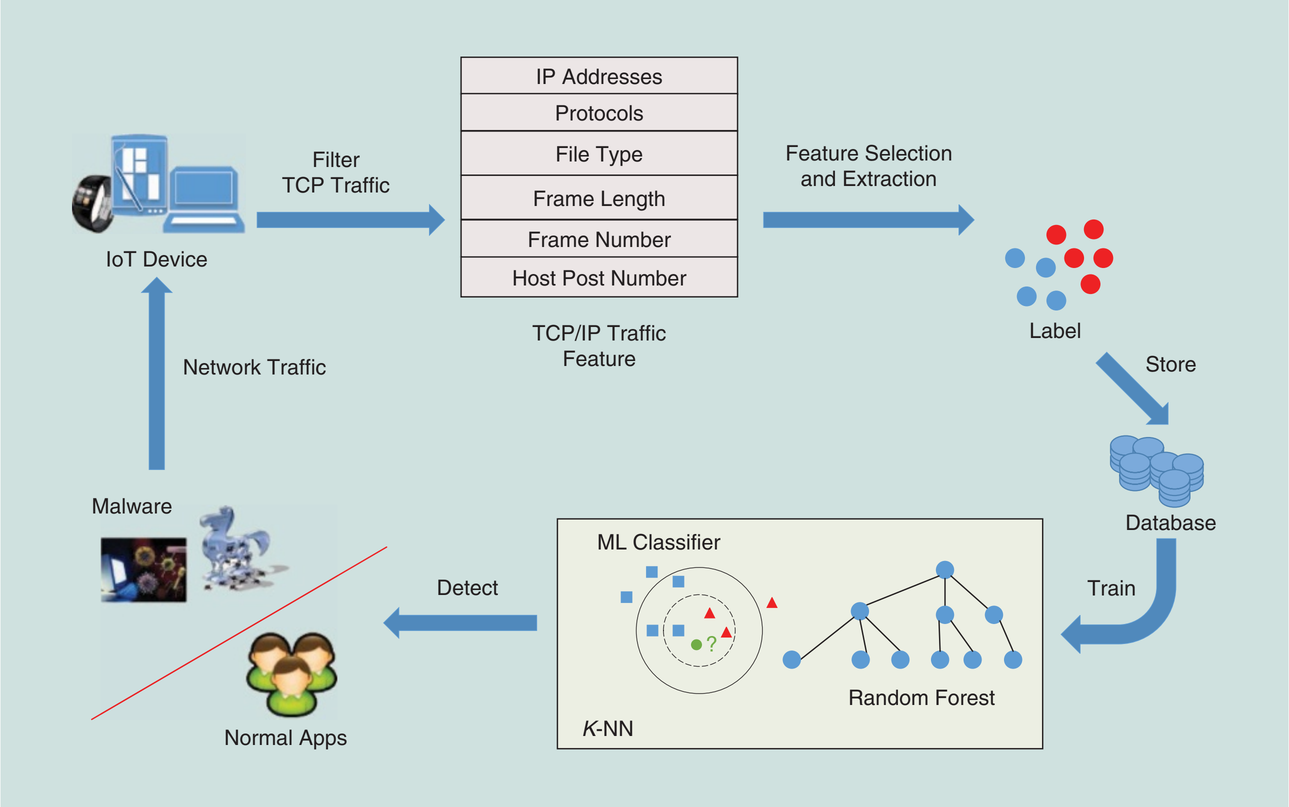


FIGURE 5. An illustration of ML-based malware detection.

IoT devices can offload app traces to the security servers at the cloud or edge devices to detect malware with a larger malware database, faster computation speed, larger memories, and more powerful security services. The optimal proportion of the app traces to offload depends on the radio channel state to each edge device and the number of the generated app traces. RL techniques can be applied for an IoT device to achieve the optimal offloading policy in a dynamic malware-detection game without being aware of the malware and app-generation models [11].

In a malware-detection scheme as developed in [11], an IoT device can apply Q-learning to achieve the optimal offloading rate without knowing the trace generation and radio bandwidth model of the neighboring IoT devices. As shown in Figure 6, the IoT device divides real-time app traces into a number of portions and observes the user density and radio channel bandwidth to formulate the current state. The IoT device estimates the detection accuracy gain, detection latency, and energy consumption to evaluate the utility received in this time slot. This scheme improves the detection accuracy by 40%, reduces the detection latency by 15%, and increases the utility of the mobile devices by 47% compared with the benchmark offloading strategy in [11] in a network consisting of 100 mobile devices.

The Dyna-Q-based malware detection scheme as presented in [11] exploits the Dyna architecture to learn from hypothetical experience and finds the optimal offloading strategy. This scheme utilizes both the real defense and virtual experiences generated by the Dyna architecture to improve the learning performance. For instance, this scheme reduces the detection latency by 30% and increases the accuracy by 18% compared with the detection with Q-learning [11].

To address the false virtual experiences of Dyna-Q, especially at the beginning of the learning process, the PDS-based malware detection scheme as developed in [11] utilizes the known radio channel model to accelerate the learning speed. This scheme applies the known information regarding the network, attack, and channel models to improve the exploration efficiency and utilizes Q-learning to study the remaining unknown state space. This scheme increases the detection accuracy by 25% compared with the Dyna-Q-based scheme in a network consisting of 200 mobile devices [11].

Conclusions and future work

In this article, we have identified IoT attack models and learning-based IoT security techniques, including IoT authentication, access control, malware detection, and secure offloading, which are shown to be promising protection for the IoT. Several challenges have to be addressed to implement the learning-based security techniques in practical IoT systems.

- *Partial state observation*: Existing RL-based security schemes assume that each learning agent knows the accurate state and evaluates the immediate reward for each action in time. In addition, the agent has to tolerate the bad strategies—especially at the beginning of the learning process. However, IoT devices usually have difficulty

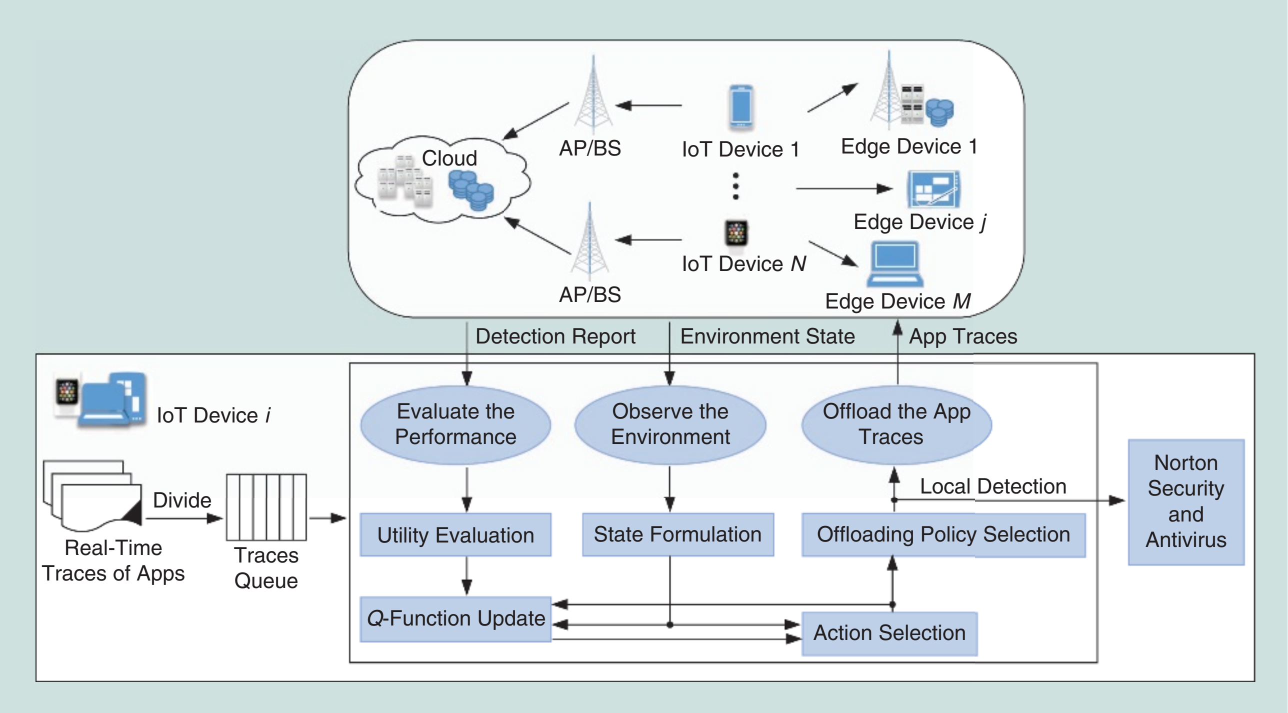


FIGURE 6. An illustration of ML-based malware detection with offloading.

estimating the network and attack state accurately and have to avoid the security disaster due to a bad policy at the beginning of the learning process. A potential solution is transfer learning [30], which explores existing defense experiences with data mining to reduce random exploration, accelerates the learning speed, and decreases the risks of choosing bad defense policies at the beginning of the learning process. In addition, backup security mechanisms have to be provided to protect IoT systems from the exploration stage in the learning process.

- *Computation and communication overhead*: Many existing ML-based security schemes have intensive computation and communication costs and require a large number of training data and a complicated feature-extraction process [9]. Therefore, new ML techniques with low computation and communication overhead such as dFW have to be investigated to enhance security for IoT systems, especially for the scenarios without cloud-based servers and edge computing.
- *Backup security solutions*: To achieve optimal strategy, the RL-based security methods have to explore the "bad" security policy that sometimes can cause network disaster for IoT systems at the beginning learning stage. The intrusion detection schemes based on unsupervised learning techniques sometimes have misdetection rates that are nonnegligible for IoT systems. Supervised and unsupervised learning sometimes fail to detect the attacks due to oversampling, insufficient training data, and bad feature extraction. Therefore, backup security solutions have to be designed and incorporated with the ML-based security schemes to provide reliable and secure IoT services.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under grants 61671396, 61671398, 61472335, and 61572538; the Fundamental Research Funds for the Central Universities under grant 17LGJC23; the open research fund of the National Mobile Communications Research Laboratory, Southeast University (2018D08); the Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China (ICT1800386); and the U.S. National Science Foundation under grants CNS-1404118, CNS-1423020, and CNS-1149611.

Authors

Liang Xiao (lxiao@xmu.edu.cn) received her B.S. degree in communication engineering from Nanjing University of Posts and Telecommunications, China, in 2000, her M.S. degree in electrical engineering from Tsinghua University, China, in 2003, and her Ph.D. degree in electrical engineering from Rutgers University, New Jersey, in 2009. She is currently a professor in the Department of Communication Engineering, Xiamen University, Fujian, China. She has been an associate editor of *IEEE Transactions on Information Forensics and Security* and *IET Communications*. Her research interests include wireless security, smart grids, and wireless communications. She won the Best Paper Award for the 2016 IEEE International Conference on Computer Communications Big Security Workshop. She has been a visiting professor with Princeton University, Virginia Tech, and the University of Maryland, College Park. She is a Senior Member of the IEEE.

Xiaoyue Wan (23320161153393@stu.xmu.edu.cn) received her B.S. degree in communication engineering from Xiamen

University, Fujian, China, in 2016, where she is currently pursuing her M.S. degree in the same field.

Xiaozhen Lu (23320170155538@stu.xmu.edu.cn) received her B.S. degree in communication engineering from Nanjing University of Posts and Telecommunications, China, in 2017. She is currently pursuing her Ph.D. degree with the Department of Communication Engineering, Xiamen University, Fujian, China. She is a Student Member of the IEEE.

Yanyong Zhang (yyzhang@winlab.rutgers.edu) received her B.S. degree in computer science from the University of Science and Technology of China, Hefei, in 1997. She a professor in the Electrical and Computer Engineering Department at Rutgers University, North Brunswick, New Jersey. She is also a member of the Wireless Information Networking Laboratory. From March to July 2009, she was a visiting scientist at Nokia Research Center, Beijing. She is the recipient of a U.S. National Science Foundation CAREER Award. She is currently an associate editor of *IEEE Transactions on Mobile Computing, IEEE Transactions on Services Computing, ACM/IEEE Transactions on Networking*, and *Elsevier Smart Health.* She has served on technical program committees of many conferences, including the IEEE International Conference on Computer Communications and the International Conference on Distributed Computing Systems. She is a Fellow of the IEEE.

Di Wu (wudi27@mail.sysu.edu.cn) received his B.S. degree from the University of Science and Technology of China, Hefei, in 2000, his M.S. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2003, and his Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong in 2007. He was a postdoctoral researcher with the Department of Computer Science and Engineering, Polytechnic Institute of New York University, Brooklyn, from 2007 to 2009, advised by Prof. K.W. Ross. He is currently a professor and the assistant dean of the School of Data and Computer Science with Sun Yat-sen University, Guangzhou, China. He was the recipient of the IEEE International Conference on Computer Communications 2009 Best Paper Award. His research interests include cloud computing, multimedia communication, Internet measurement, and network security.

References

[1] X. Li, R. Lu, X. Liang, and X. Shen, "Smart community: An Internet of things application," *IEEE Commun. Mag.*, vol. 49, no. 11, pp. 68–75, Nov. 2011.

[2] B. Firner, R. S. Moore, R. Howard, R. P. Martin, and Y. Zhang, "Poster: Smart buildings, sensor networks, and the Internet of things," in *Proc. ACM Conf. Embedded Networked Sensor Systems*, Nov. 2011, pp. 337–338.

[3] Z. Sheng, S. Yang, Y. Yu, and A. Vasilakos, "A survey on the IETF protocol suite for the Internet of things: Standards, challenges, and opportunities," *IEEE Wireless Commun.*, vol. 20, no. 6, pp. 91–98, Dec. 2013.

[4] I. Andrea, C. Chrysostomou, and G. Hadjichristofi, "Internet of things: Security vulnerabilities and challenges," in *Proc. IEEE Symp. Computers and Communication*, Larnaca, Cyprus, Feb. 2015, pp. 180–187.

[5] R. Roman, J. Zhou, and J. Lopez, "On the features and challenges of security and privacy in distributed Internet of things," *Comput. Netw.*, vol. 57, no. 10, pp. 2266–2279, July 2013.

[6] S. Chen, H. Xu, D. Liu, and B. Hu, "A vision of IoT: Applications, challenges, and opportunities with china perspective," *IEEE Internet Things J.*, vol. 1, no. 4, pp. 349–359, July 2014.

[7] J. Zhou, Z. Cao, X. Dong, and A. V. Vasilakos, "Security and privacy for cloud-based IoT: Challenges," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 26–33, Jan. 2017.

[8] L. Xiao, Y. Li, G. Han, G. Liu, and W. Zhuang, "PHY-layer spoofing detection with reinforcement learning in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10037–10047, Dec. 2016.

[9] M. Abu Alsheikh, S. Lin, D. Niyato, and H. P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," *IEEE Commun. Surveys Tutorials*, vol. 16, no. 4, pp. 1996–2018, Apr. 2014.

[10] L. Xiao, C. Xie, T. Chen, and H. Dai, "A mobile offloading game against smart attacks," *IEEE Access*, vol. 4, pp. 2281–2291, May 2016.

[11] L. Xiao, Y. Li, X. Huang, and X. J. Du, "Cloud-based malware detection game for mobile devices with offloading," *IEEE Trans. Mobile Comput.*, vol. 16, no. 10, pp. 2742–2750, Oct. 2017.

[12] M. Ozay, I. Esnaola, F. T. Yarman Vural, S. R. Kulkarni, and H. V. Poor, "Machine learning methods for attack detection in the smart grid," *IEEE Trans. Neural Networks and Learning Syst.*, vol. 27, no. 8, pp. 1773–1786, Mar. 2015.

[13] J. W. Branch, C. Giannella, B. Szymanski, R. Wolff, and H. Kargupta, "In-network outlier detection in wireless sensor networks," *Knowl. Inform. Syst.*, vol. 34, no. 1, pp. 23–54, Jan. 2013.

[14] F. A. Narudin, A. Feizollah, N. B. Anuar, and A. Gani, "Evaluation of machine learning classifiers for mobile malware detection," *Soft Comput.*, vol. 20, no. 1, pp. 343–357, Jan. 2016.

[15] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tutorials*, vol. 18, no. 2, pp. 1153–1176, Oct. 2015.

[16] R. V. Kulkarni and G. K. Venayagamoorthy, "Neural network based secure media access control protocol for wireless sensor networks," in *Proc. Int. Joint Conf. Neural Networks*, Atlanta, GA, June 2009, pp. 3437–3444.

[17] Z. Tan, A. Jamdagni, X. He, P. Nanda, and R. P. Liu, "A system for Denial-of-Service attack detection based on multivariate correlation analysis," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 2, pp. 447–456, May 2013.

[18] L. Xiao, Q. Yan, W. Lou, G. Chen, and Y. T. Hou, "Proximity-based security techniques for mobile users in wireless networks," *IEEE Trans. Inform. Forensics Security*, vol. 8, no. 12, pp. 2089–2100, Oct. 2013.

[19] Y. Gwon, S. Dastangoo, C. Fossa, and H. Kung, "Competing mobile network game: Embracing anti-jamming and jamming strategies with reinforcement learning," in *Proc. IEEE Conf. Communication and Network Security*, National Harbor, MD, Oct. 2013, pp. 28–36.

[20] M. A. Aref, S. K. Jayaweera, and S. Machuzak, "Multi-agent reinforcement learning based cognitive anti-jamming," in *Proc. IEEE Wireless Communication and Networking Conf.*, San Francisco, CA, Mar. 2017, pp. 1–6.

[21] Y. Li, D. E. Quevedo, S. Dey, and L. Shi, "SINR-based DoS attack on remote state estimation: A game-theoretic approach," *IEEE Trans. Contr. Network Syst.*, vol. 4, no. 3, pp. 632–642, Apr. 2016.

[22] G. Han, L. Xiao, and H. V. Poor, "Two-dimensional anti-jamming communication based on deep reinforcement learning," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, New Orleans, LA, Mar. 2017, pp. 2087–2091.

[23] C. Shi, J. Liu, H. Liu, and Y. Chen, "Smart user authentication through actuation of daily activities leveraging WiFi-enabled IoT," in *Proc. ACM Int Symp. Mobile AdHoc Networking and Computing*, Chennai, India, July 2017, pp. 1–10.

[24] X. He, H. Dai, and P. Ning, "Improving learning and adaptation in security games by exploiting information asymmetry," in *Proc. IEEE Conf. Computer Communication*, Hongkong, China, May 2015, pp. 1787–1795.

[25] V. Mnih, K. Kavukcuoglu, D. Silver, et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Jan. 2015.

[26] Z. Yan, P. Zhang, and A. V. Vasilakos, "A survey on trust management for Internet of things," *J. Netw. Comput. Appl.*, vol. 42, no. 3, pp. 120–134, June 2014.

[27] L. Xiao, X. Wan, and Z. Han, "PHY-layer authentication with multiple landmarks with reduced overhead," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1676–1687, Mar. 2018.

[28] J. Yu, H. Lee, M. S. Kim, and D. Park, "Traffic flooding attack detection with SNMP MIB using SVM," *Comput.Commun.*, vol. 31, no. 17, pp. 4212–4219, Oct. 2008.

[29] R. Roman, J. Lopez, and M. Mambo, "Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges," *Future Gener. Comput. Syst.*, vol. 78, no. 3, pp. 680–698, Jan. 2018.

[30] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowledge Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

Jiangfan Zhang, Rick S. Blum, and H. Vincent Poor

Approaches to Secure Inference in the Internet of Things

Performance bounds, algorithms, and effective attacks on IoT sensor networks

INTERNET OF THINGS—ISTOCKPHOTO.COM/IAREMENKO
CIRCUITS—IMAGE LICENSED BY INGRAM PUBLISHING

The Internet of Things (IoT) improves pervasive sensing and control capabilities via the aid of modern digital communication, signal processing, and massive deployment of sensors but presents severe security challenges. Attackers can modify the data entering or communicated from the IoT sensors, which can have a serious impact on any algorithm using these data for inference. This article describes how to provide tight bounds (with sufficient data) on the performance of the best unbiased algorithms estimating a parameter from the attacked data and communications under any assumed statistical model describing how the sensor data depends on the parameter before attack. The results hold regardless of the unbiased estimation algorithm adopted, which could employ deep learning, machine learning, statistical signal processing, or any other approach. Example algorithms that achieve performance close to these bounds are illustrated. Attacks that make the attacked data useless for reducing these bounds are also described. These attacks provide a guaranteed attack performance in terms of the bounds regardless of the algorithms the unbiased estimation system employs. References are supplied that provide various extensions to all of the specific results presented in this article and a brief discussion of low-complexity encryption and physical layer security is provided.

Introduction

The IoT will introduce an unprecedented increase in sensor resources and data-producing sensor-like objects for many applications. Over 1 trillion IoT sensors, machines, objects, and devices are expected to be connected to the Internet by 2022. The top three IoT applications by market share are anticipated to be health care (41%), manufacturing, (37%), and electricity grids (7%). Even more impressive, IoT smart objects are expected to generate 45% of all Internet traffic by 2022. While the Internet has been available for many years, the integration of sensing technology into the Internet is still very immature and brings new problems that have not yet been addressed. Serious security concerns for IoT systems have already been demonstrated, and the future brings even more concerns. For

Digital Object Identifier 10.1109/MSP.2018.2842261
Date of publication: 28 August 2018

1053-5888/18©2018IEEE

example, self-driving cars could become dangerous weapons unless adequate security solutions are developed. For these reasons, many researchers are focused on finding new cybersecurity technologies for the IoT to augment current technology [1], [2]. Each new technology, including the inferential sensor processing technology that is the focus of this article, can form one layer of a multilayer security paradigm, with the other layers employing different approaches drawn from both new and existing alternatives. The hope is that if one layer is defeated, the other layers could still provide protection.

Typically, large IoT systems are composed of low-cost and spatially distributed sensor nodes with limited battery power and low computing capacity, which makes them particularly vulnerable to cyberattacks by adversaries. This has led to great interest in studying the vulnerability of the IoT in various applications and from different perspectives; see [3]–[24] and the references therein. Moreover, due to the dominance of digital technology, quantization has been widely employed at the sensors in IoT systems. The more recent topic of cybersecurity for IoT has received less attention than the topic of cybersecurity for other systems, but the increasing adoption of sensors and IoT networks makes this a very important issue. This article focuses on machine-learning (we do not study attacks on the training here) and signal processing approaches to the development of security in inferential sensor processing for the IoT using quantized data. The discussion will mainly focus on estimation in the presence of active cyberattacks that manipulate the data in IoT systems, although the ideas can also be generalized in many interesting ways beyond estimation applications. To provide a clear picture in a limited space, we focus on techniques to allow the estimation system to identify such attacks and perform robust processing in their presence. In fact, we provide tight bounds (for sufficient sample sizes) on the best possible performance the unbiased estimation system can achieve. We also describe optimized attacks from the attacker's point of view. At the end of the article, we provide a brief discussion of some specific aspects of some related topics of interest, including eavesdropping, secrecy, encryption, and authentication.

The topic of impact and mitigation of cyberattacks on systems solving hypothesis testing problems was studied in [3], [4], [7]–[9], and references therein. Investigations on cyberattacks on estimation systems have been studied in [6], [7], [10]–[15], [17], [23], and references therein. The early work in [3], [4], [6], [7]–[9], [13]–[15], and [23] set the tone for many later investigations and influenced most of the discussions in this article. In particular, the impact and mitigation of cyberattacks on systems solving hypothesis testing problems was studied in [3], [4], [7], [9], and references therein. Distributed detection in tree topologies in the presence of cyberattacks was considered in [8]. Investigations of cyberattacks on estimation systems have been studied in [6], [7], [10]–[15], [17], [23], and references therein. The problem of distributed spectrum sensing in a cognitive radio network under cyberattacks was studied in [4], [5], and [25]. Several cyberattack detection techniques were proposed for IoT localization systems in [6], [12], [18], and [19]. More recently, the data-injection attacks in smart grids were considered in [13]–[15], [26], and the references therein.

According to where they occur, cyberattacks in IoT systems can be categorized into two classes, as illustrated in Figure 1. We call any attack modifying a signal in the IoT system prior to quantization a *spoofing attack*. It has been shown [12] that the same changes in the signals in the IoT system produced by any spoofing attack can also be produced by changing the data going into the sensors to be different from that coming from the physical phenomenon being monitored. We call any attack modifying a signal in the IoT system after quantization a *man-in-the-middle attack* (*MiMA*). The same changes in the signals in the IoT system produced by any MiMA [12] can also be produced by changing the quantized data transmitted

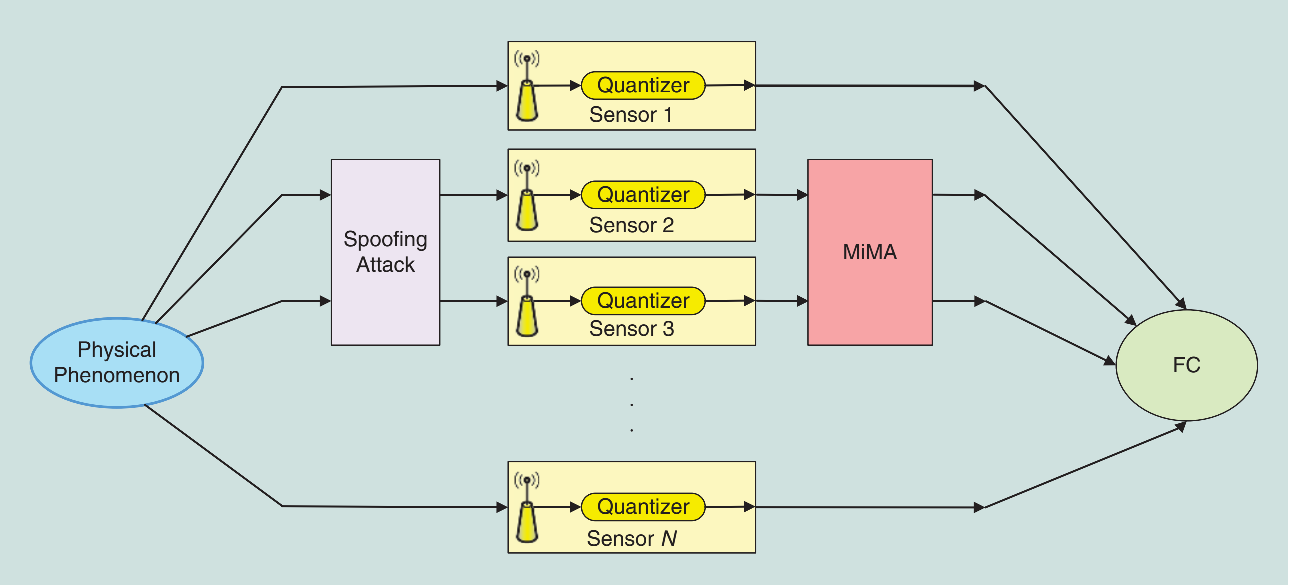


FIGURE 1. Cyberattacks in IoT systems.

by the sensors. Further, combinations of these two possible classes of attacks can represent any type of possible attack even if the actual attack modifies the sensor hardware or software as opposed to changing the data entering or leaving the sensor node.

MiMAs caused by an attacker intercepting a communication packet and changing its contents or by an attacker forcing a sensor node to transmit false data have received previous research attention. For instance, MiMAs were studied for distributed spectrum sensing in a cognitive radio network in [4], [5], and [25]. The distributed detection problem in the presence of MiMAs was investigated in [3]. Mitigation techniques for MiMAs were studied in localization problems in [6] and [12]. Spoofing attacks have also been studied for localization problems; see [18], [19], and the references therein. In [18, Table I], a summary of different types of spoofing attacks for localization problems is provided. The dangers of spoofing attacks on global positioning system (GPS) receivers that provide important information to everything from car navigation to national power grids have drawn serious public concern [27], [28]. Radar and sonar systems also suffer from spoofing attacks in practice. As an example of a spoofing attack technique, the application of an electronic countermeasure (ECM), which is designed to deceive a radar or sonar system, can critically degrade the detection and estimation performance of the system [29]. One popular technique for the implementation of ECMs employs digital radio-frequency memory (DRFM) to store a received radar signal and transmit it back to the radar receiver to confuse the victim radar system. DRFM can mislead the estimation of the range of the target by altering the delay of the pulses received by the radar system and fool the system into incorrectly estimating the velocity of the target by introducing a fake Doppler shift in the retransmitted signal [30]. Since radar systems are being installed by most car manufacturers, with the ultimate application being self-driving cars, spoofing attacks are potentially very dangerous. The data-injection attack in smart grids is another typical example of a spoofing attack; see [13]–[15], [26], and the references therein.

Regarding MiMAs, we focus on the fundamental problems in identifying and mitigating the impact of malicious cyberattacks. In particular, it is shown that, under some assumptions, it is possible to correctly identify the attacked sensors and categorize them into differently attacked groups. One such assumption is that the largest group of similarly attacked sensors are unattacked. Furthermore, once the differently attacked sensors have been categorized, necessary and sufficient conditions are provided that describe when the attacked sensor data can and cannot improve the estimation performance in terms of the Cramér–Rao bound (CRB).

All existing research on attacks on IoT systems performing inference considers cases in which the attacker replaces the unattacked sensor or communication data by a function of the unattacked data where the form of the function is known down to some unknown scalar quantities which are called *attack parameters*. For example, a specific type of spoofing attack, called a *data-injection attack*, adds an unknown attack parameter to the sensor data. Thus, the function here is a linear function with unit slope, and the attack parameter is the value added to the sensor data. We consider much more general types of attacks and describe the functions that guarantee the IoT estimation system can achieve, at best, a given level of performance no matter what approach the estimator takes. This shows the existence of very powerful attacks, from the attacker's point of view, such that the attacker is guaranteed to force the estimation system to have performance below some unacceptable value. To be precise, for a generalized spoofing attack using known functions with unknown attack parameters, necessary and sufficient conditions are provided under which the attack provides a guaranteed attack performance in terms of CRB degradation regardless of the processing the IoT system employs, thus defining a highly desirable attack. Further analysis of these attacks reveals that the quantization imposes a limit on the capability of the system to defend against attacks, which can be exploited to construct an optimal attack by properly employing a sufficiently large dimensional attack vector parameter relative to the number of quantization symbols employed.

The most general attacks, which include combinations of MiMAs and spoofing attacks, are illustrated in the section "General Attacks in Vector Parameter Estimation Systems," when estimating the location of an object. With the help of two secure sensors, a class of detectors is proposed to detect the attacked sensors by scrutinizing the existence of a geometric inconsistency. Moreover, it is shown that the error probability of the proposed attack detector decays exponentially by employing large deviations techniques.

Originally motivated by our research on cybersecurity, we reveal a fundamental limitation on quantized estimation systems not under attack. A critical quantity called *inestimable dimension for quantized data* (*IDQD*) is introduced, which does not depend on the estimation problem, the quantization regions, or the exact statistical models of the observations but instead depends only on the number of sensors and on the precision of the quantizers employed by the system. It is shown that, if the dimension of the desired vector parameter is larger than the IDQD of the quantized estimation system, then the Fisher information matrix (FIM) for estimating the desired vector parameter is singular, and, moreover, there exist infinitely many nonidentifiable vector parameter points in the vector parameter space.

MiMAs

To introduce a simple problem, we consider a set of N distributed IoT sensors, each making K time observations of a deterministic scalar parameter θ corrupted by additive noise. At the jth sensor, the observation at the kth time instant is described by

$$x_{jk} = \theta + n_{jk}, \forall j = 1,2,...,N, \forall k = 1,2,...,K, \tag{1}$$

where n_{jk} denotes an additive noise sample with zero-mean probability density function (pdf) $f(n_{jk})$ and $\{n_{jk}\}$ is an independent and identically distributed sequence. (Extensions to general estimation problems and nonbinary quatization are considered in [11].) Each observation x_{jk} is individually quantized, and the result is denoted by u_{jk}. All of the quantized observations are sent to the fusion center (FC) for use in estimating θ. While we allow these communications to be attacked, we ignore any other errors in the communications to keep things simple, including those due to noise or fading.

Lately, there has been great interest in the extreme case where each sensor is restricted to transmitting a single bit per observation to the FC. A basic approach is to decide $u_{jk} = 1$ if $x_{jk} > \nu$, where ν is a fixed threshold, and $u_{jk} = 0$ otherwise. Thus, without attacks $\Pr(u_{jk} = 0 \mid \theta) = F(\nu - \theta)$ and $\Pr(u_{jk} = 1 \mid \theta) = 1 - F(\nu - \theta)$, where $F(x) \overset{\Delta}{=} \int_{-\infty}^{x} f(t)dt$ denotes the cumulative distribution function (cdf) corresponding to the pdf $f(x)$. By employing the invariance of the maximum likelihood estimate (MLE), the naive MLE (NMLE), the MLE formulated under the assumption of no attack, of the parameter θ can be expressed as

$$\hat{\theta}_{\mathrm{NML}} = \nu - F^{-1}\left(1 - \frac{1}{KN}\sum_{j=1}^{N}\sum_{k=1}^{K} u_{jk}\right), \tag{2}$$

which, without the presence of an adversary, can be expected to provide asymptotically unbiased and efficient estimation.

Let, at most, P distinct malicious attacks (P arbitrary) be launched at a given time, where each attack follows a fairly general adversary model to be described next. Let $\mathcal{A}_p$ denote the set of sensors subjected to the pth attack, and let $\tilde{u}_{jk}$ represent the after-attack quantized observation, which is a modified version of u_{jk}. The statistical description of the pth attack can be represented by a probability transition matrix $\boldsymbol{\Psi}_p$,

$$\boldsymbol{\Psi}_p \overset{\Delta}{=} \begin{bmatrix} \psi_{p,0} & 1 - \psi_{p,1} \\ 1 - \psi_{p,0} & \psi_{p,1} \end{bmatrix}, \tag{3}$$

where $\psi_{p,0} \overset{\Delta}{=} \Pr(\tilde{u}_{jk} = 0 \mid u_{jk} = 0)$ and $\psi_{p,1} \overset{\Delta}{=} \Pr(\tilde{u}_{jk} = 1 \mid u_{jk} = 1)$ are attack parameters that determine the modification probabilities (flipping probabilities). Here, we assume the attacker does not know θ, and so the attack parameters do not depend on θ. Extended discussion of various cases in which the attacker has more or less information about the estimation system and the estimation problem are considered in [11]. Due to the pth attack, the after-attack probability mass function (pmf) of the observations can be related to the before-attack pmf using

$$\begin{bmatrix} 1 - \tilde{p}(\boldsymbol{\Psi}_p, \theta) \\ \tilde{p}(\boldsymbol{\Psi}_p, \theta) \end{bmatrix} \overset{\Delta}{=} \begin{bmatrix} \Pr(\tilde{u}_{jk} = 0 \mid \theta) \\ \Pr(\tilde{u}_{jk} = 1 \mid \theta) \end{bmatrix} = \boldsymbol{\Psi}_p \begin{bmatrix} \Pr(u_{jk} = 0 \mid \theta) \\ \Pr(u_{jk} = 1 \mid \theta) \end{bmatrix}. \tag{4}$$

For the sake of expressing the after-attack pmfs of observations in a uniform form for both attacked and unattacked sensors, define the set $\mathcal{A}_0$ of unattacked sensors, that is, if $j \in \mathcal{A}_0$, then $\tilde{u}_{jk}$ and u_{jk} have the same distribution.

Assumption 1

The following assumption on attacks is made throughout this article.

1) Over the K sample estimation time interval described in (1) and for all p, the pth attack is statistically described as in (4) for all the sensors in the set $\mathcal{A}_p$. The set $\mathcal{A}_p$ and the attack parameters are unknown to the FC. Let $\mathcal{P}_p \overset{\Delta}{=} |\mathcal{A}_p|/N$. Moreover, we assume that the group of unattacked sensors is the largest group $\mathcal{P}_0 > \mathcal{P}_p + \Delta_0$ for all $p \geq 1$ where Δ_0 is a positive constant. Further, the sets $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_P$ are disjoint $\mathcal{A}_p \cap \mathcal{A}_{p'} = \emptyset \quad if \quad p \neq p'$.
2) Significant attacks. Since attacks that cause very small changes to $\tilde{p}(\boldsymbol{\Psi}_0, \theta)$ cause very little impact on performance (similar to small noise), we only consider attacks that produce at least a minimum distortion d_{impact} on $\tilde{p}(\boldsymbol{\Psi}_0, \theta)$ and tamper with at least Δ percent of sensors so that

$$|\tilde{p}(\boldsymbol{\Psi}_p, \theta) - \tilde{p}(\boldsymbol{\Psi}_0, \theta)| \geq d_{\mathrm{impact}}, \quad \forall p = 1, 2, \ldots, P, \tag{5}$$

$$\mathcal{P}_p \overset{\Delta}{=} |\mathcal{A}_p|/N \geq \Delta > 0, \quad \forall p = 1, 2, \ldots, P. \tag{6}$$

3) Various attacks. The changes caused by two distinct types of attacks are considerably different; otherwise, these two types of attacks can be treated as identical. To this end, we assume that

$$|\tilde{p}(\boldsymbol{\Psi}_l, \theta) - \tilde{p}(\boldsymbol{\Psi}_m, \theta)| \geq d_{\mathrm{diff}}, \quad \forall l \neq m. \tag{7}$$

It is worth mentioning that the adversary model assumed in (4) can change the after-attack pmf to have any desired valid values satisfying (5) and (7) through proper choice of the two attack parameters $\psi_{p,0}$ and $\psi_{p,1}$. In this sense, it is a fairly general adversary model.

Identification and categorization of attacked sensors

The following theorem describes the identification and categorization of the attacked sensors.

Theorem 1

Under Assumption 1, for any N as $K \to \infty$, the FC can always identify from the observations, without further knowledge, a $\mathcal{P}_0$ percentage group of sensors that contains 0% attacked sensors. Similarly, as $K \to \infty$, the FC is also able to identify P other groups of sensors that, respectively, make up $\{\mathcal{P}_p\}_{p=1}^{P}$ percent of all sensors, such that for $p = 1, 2, \ldots, P$, group p contains 0% sensors not experiencing attack p.

On the other hand, assume each sensor observes a finite number K of time samples such that

$$K \geq -\frac{8 \ln 2}{\gamma^{*} \min\{\Delta\Delta_0, \Delta^2\}} + 1, \tag{8}$$

where γ^{*} is a constant defined in [10]. Under Assumption 1, as $N \to \infty$, the FC can determine P and a group of sensors $\tilde{\mathcal{A}}_p$ corresponding to $\mathcal{A}_p$, for $p = 1, \ldots, P$, with $\tilde{\mathcal{P}}_p \overset{\Delta}{=} |\tilde{\mathcal{A}}_p|/N, \mathcal{P}_p^{*} \overset{\Delta}{=} |(\tilde{\mathcal{A}}_p \backslash \mathcal{A}_p) \cup (\mathcal{A}_p \backslash \tilde{\mathcal{A}}_p)|/N$, and $\delta \overset{\Delta}{=} -(4 \ln 2 / \Delta (K-1) \gamma^{*})$, which satisfy

$$0 \leq |\tilde{\mathcal{P}}_p - \mathcal{P}_p| \leq \mathcal{P}_p^{*} < \delta. \tag{9}$$

One should notice the stark differences in Theorem 1 when we increase N to large values instead of K. Given that we define a sensor as attacked or unattacked, this does make sense. When we are given more data at a given sensor for which we already had some data, then the new data will certainly help us better categorize the statistical model for the data at this sensor. If we increase K at a group of fixed sensors, this will help us determine which sensors are attacked, which are not, and which sensors are similarly attacked. If we are given data from a new sensor, from which we had not previously been given data, then we are also given a new problem: "Is this sensor attacked?" Thus, given our problem formulation, increasing K to large values is more helpful than increasing N to large values.

The essential idea toward accomplishing the identification and categorization of the attacked sensors is to recognize that the statistical description of the data at the differently attacked sensors will be significantly different based on Assumption 1. Thus, one could estimate the pmfs of the quantized data at each sensor using histograms and then classify the sensors into different groups representing the different attacks or the group of unattacked sensors. As the number of observations at each sensor K becomes large, it seems reasonable that the estimates become more accurate for larger K. Many other methods can also be used for identification and categorization. We can use the estimate in (2) to also see the statistical differences from sensor to sensor for sufficiently large K given the good properties of this estimate for the described problem. Note that Assumption 1 defines significant differences in the pmfs of unattacked and attacked data [numerical values for d_{impact} in (5) can be chosen based on which differences cause significant performance degradation to the estimation performance when the data are assumed to be unattacked]. Note that Assumption 1 also defines significant differences in the pmfs of differently attacked data [numerical values for d_{diff} in (7) can be chosen based on which differences cause significant degradation if ignored]. We also need a way to distinguish which group is unattacked. If we know the largest group is unattacked, as assumed in Assumption 1, or if we have some protected sensors, these are some methods to distinguish which group is unattacked from among the groups of sensors deemed to be statistically different.

Estimation performance improvement via using attacked sensor data

As demonstrated by Theorem 1, when each sensor accumulates sufficiently many time samples, then the FC is able to determine the number of attacks in the network and very accurately categorize the sensors into different groups according to distinct types of attacks. In the rest of this section, we assume that the sensors have been well categorized into the groups $\{\mathcal{A}_p\}_{p=0}^{P}$, and we attempt to estimate the desired parameter θ. For simplicity, we assume the categorizations are exactly correct $(K \to \infty)$, but the following results would only be approximately true if errors are made $(K \not\to \infty)$. There are two approaches:

1) Ignore the data at the attacked sensors and just employ the data at the unattacked sensors to estimate the desired parameter. We refer to this approach as the *simple estimation approach* (*SEA*).
2) Use the data at the attacked sensors and jointly estimate the desired parameter and the unknown attack parameters.

It requires less complexity to take approach 1), which avoids estimating any parameters describing the attacks. However, to attempt to take approach 2), and potentially do better than approach 1), we will investigate the performance of the joint estimation of the desired parameter and the unknown attack parameters. Let $\boldsymbol{\theta} \overset{\Delta}{=} [\theta, \psi_{1,0}, \psi_{1,1}, ..., \psi_{P,0}, \psi_{P,1}]^T$ denote a vector containing the desired scalar parameter θ along with all of the unknown parameters of the attacks. The estimation performance is evaluated by the mean-squared error (MSE), which is lower bounded in a positive definite sense using

$$\mathbb{E}\{[\hat{\boldsymbol{\theta}}(\mathbf{u}) - \boldsymbol{\theta}][\hat{\boldsymbol{\theta}}(\mathbf{u}) - \boldsymbol{\theta}]^T\} \succeq \mathbf{J}^{-1}(\boldsymbol{\theta}), \tag{10}$$

where $\hat{\boldsymbol{\theta}}$ is any unbiased estimator of $\boldsymbol{\theta}$, $\mathbf{u}$ denotes the vector that contains all employed quantized observations $\{u_{jk}\}$, $\mathbf{J}(\boldsymbol{\theta})$is the FIM, and the (1, 1) component of $\mathbf{J}^{-1}(\boldsymbol{\theta})$ is the CRB for estimating the desired scalar parameter θ. Note that the CRB is an asymptotically achievable bound on MSE. In typical applications, a good estimator with the required number of observations to achieve the desired performance usually performs close to the CRB. We will make use of the CRB and FIM to benchmark the estimation performance of unbiased parameter estimators. In our studies of inference for the IoT in this article, we restrict attention to unbiased estimators. Extensions to biased estimators is a topic of current research. If the FIM is singular for the data from a specific sensor, then those data are no longer useful to reduce the MSE when the data are fused with data from other sensors [17]. An attacker can create this situation with a proper attack [17]. Thus, the FIM can provide a rigorous way to identify good attacks that make the attacked data useless. Knowing that attacked data are useless for reducing the MSE when those data are fused with data from other sensors is also useful in the estimation procedure [10]. Thus, the CRB and FIM are very powerful while being relatively easy to compute. General calculations of MSE are generally intractable. This explains why the CRB is the most widely used lower bound and why analysis based on the FIM is so common. One can certainly expand the work discussed here to go beyond these metrics, but there will be a cost in terms of computational complexity and the simplicity of explanation obtained by simple closed-form expressions.

It is shown in [10] that using the fixed threshold approach described before (2) will not allow joint estimation of the desired and attack parameters, since the FIM for that estimation is singular. This phenomenon is explained by the theory we provide in the section "Implications for Unattacked Systems." There we show that the quantized observations from a given quantization approach are really only capable of accurately estimating a parameter with dimension smaller than a given value. The quantization approach with a common threshold for all sensors and for all samples at each sensor can only estimate a scalar parameter for the given problem. This approach cannot jointly estimate both the desired scalar parameter and the attack parameters. To overcome this, we can employ a quantization scheme that allows us to estimate a larger dimensional parameter, with a dimension $2P+1$ for the $2P$ attack parameters and the desired scalar parameter θ. In particular, we define a set of Q distinct thresholds $\mathfrak{T} = \{\nu_1, \nu_2, ..., \nu_Q\}$ and employ different thresholds over Q distinct time slots $\{\mathcal{T}_t\}_{t=1}^{Q}$, while using the same threshold at each sensor. We refer to this approach as the *time-variant quantization approach* (*TQA*). Let $\tilde{p}_p^{(t)} \triangleq \Pr(\tilde{u}_{jk} = 1 | \boldsymbol{\theta})$ for $j \in \mathcal{A}_p, k \in \mathcal{T}_t$, and let $\boldsymbol{\Xi}_p \triangleq \frac{d}{d\theta}[\tilde{p}_p^{(1)}, \tilde{p}_p^{(2)}, \ldots, \tilde{p}_p^{(Q)}]$.

In Theorem 2, we provide necessary and sufficient conditions under which the CRB performance of estimating the desired scalar parameter θ can be improved by employing observations from an attacked sensor.

Theorem 2

The FIM for estimating $\boldsymbol{\theta}$ is nonsingular provided that $Q \geq 2$. Moreover, the CRB for the desired scalar parameter θ can be improved by utilizing the observations from the set of attacked sensors in our proposed fashion (TQA) if and only if for some $p \in \{1, 2, \ldots, P\}$, $\operatorname{rank}(\boldsymbol{\Xi}_p) = 3$. Otherwise, there is no CRB improvement, but also no loss in CRB, from utilizing the attacked observations.

In particular, by employing the TQA, the relative CRB gain from utilizing the observations at the attacked sensors is

$$\frac{\text{CRB using SEA}}{\text{CRB using TQA}} = 1 + \frac{1}{[\boldsymbol{\Gamma}_0]_{1,1}} \sum_{p=1}^{P} \frac{\det(\boldsymbol{\Gamma}_p(\{1, 2p, 2p+1\}, \{1, 2p, 2p+1\}))}{\det(\boldsymbol{\Gamma}_p(\{2p, 2p+1\}, \{2p, 2p+1\}))}, \tag{11}$$

where $\boldsymbol{\Gamma}_p(\{i_1, i_2, \ldots, i_L\}, \{j_1, j_2, \ldots, j_M\})$ denotes the submatrix of $\boldsymbol{\Gamma}_p \triangleq \boldsymbol{\Xi}_p \boldsymbol{\Lambda}_p \boldsymbol{\Xi}_p^T$ (p = 0, 1, …, P), which consists of the elements located in the $\{i_l\}_{l=1}^{L}$th rows and $\{j_m\}_{m=1}^{M}$th columns. $[\boldsymbol{\Gamma}_0]_{1,1}$ is the (1,1) component of $\boldsymbol{\Gamma}_0$. The matrix $\boldsymbol{\Lambda}_p$ is a Q-by-Q diagonal matrix, and the tth diagonal element of $\boldsymbol{\Lambda}_p$ is $K_t \mathcal{P}_p / \tilde{p}_p^{(t)}(1 - \tilde{p}_p^{(t)})$, where K_t is the number of time samples in $\mathcal{T}_t$.

Interpretation of Theorem 2 and (11) is now given. Recall that the CRB is a lower bound on the MSE of any unbiased estimator. The CRB is achieveable with a reasonable number of observations. The ratio of the CRB of the approach that ignores the attacked data to the CRB of the approach using the attacked data is shown in (11). Here, we see the power of the CRB in allowing us to obtain fairly simple closed-form expressions that we could not obtain using general expressions of MSE. One of the most interesting aspects of (11) is when the ratio is larger than unity. If the ratio is larger than unity, then it is advantageous to use the attacked data in terms of CRB. If the ratio is unity, then the estimator can ignore the attacked data. From (11), and noting the provable nonnegativity of the second term due to the positive semidefiniteness of $\boldsymbol{\Gamma}_p$, the ratio must be unity or larger. Thus, (11) describes the utility of the attacked data in a very simple manner. Note that (11) also describes the exact value of the improvement. Since the determinant of any rank deficient matrix is zero, (11) also verifies Theorem 2, since the denominator matrix in the second term is always full rank and the numerator needs the matrix $\boldsymbol{\Xi}_p$ referred to in Theorem 2 to have rank three for some p to ensure that the ratio of CRBs will be greater than unity. Note that each entry of one of the columns of $\boldsymbol{\Xi}_p$ is obtained by taking a derivative with respect to one of the components of the vector $\boldsymbol{\theta}$. Since the pmf of the data under the pth attack can depend only on θ and the two pth attack parameters and not on the other attack parameters, then $\boldsymbol{\Xi}_p$ can have at most three nonzero rows. Due to this, $\boldsymbol{\Gamma}_p$ can have only nine nonzero entries, which explains the form of (11).

Generalizations and motivating IoT estimation problems

In [11], we provide extensions to the previously discussed results for nonbinary quantization and general estimation problems. For these cases, we provide a theorem similar to Theorem 1 on the ability to categorize and classify the differently attacked and unattacked sensors. After classification, one can similarly judge if the data at a group of similarly attacked sensors can be useful to improve estimation performance in terms of CRB. Once again, some attacks will make the attacked data useless for this purpose. This generalization allows us to consider many important IoT estimation applications.

One application that has drawn significant attention lately is that of self-driving cars. Attacks in this application are especially concerning since loss of life could result. This application clearly convinces us of the importance of further developing the kind of theory initiated in this article. It turns out that most car manufacturers are convinced that the best way to stop self-driving cars from injuring people is to fuse radar and video data. In fact, some may want to fuse other sensors as well. Car manufacturers are all developing inexpensive integrated circuit chips to fully incorporate the radar processing. Interestingly, when these inexpensive integrated circuit chips become available, this will encourage extensive use of radar in all kinds of applications and products, beyond autonomous vehicles. Surveillance applications will certainly benefit. In fact, the same process will likely be followed for other complicated sensors. Thus, when inexpensive integrated circuits become available for these other sensors, this will encourage extensive use of these sensors in all kinds of applications and products. Since these sensors can be attacked, methods for protecting these sensors, like the ones presented here, become extremely important. Attacks on the sensors (the GPS is also a sensor) or communications in self-driving cars are one application motivating this work.

In [12], we focus on location estimation under possible simultaneous MiMAs and spoofing attacks, but similar approaches can be applied for other vector parameter estimation problems. These vector parameter estimation problems can be important in medical, manufacturing, and smart grid applications, among others. We discuss [12] in more detail in the section "General Attacks in Vector Parameter Estimation Systems."

Illustrative example: Identification and categorization of attacked sensors

Consider a network with $N = 10$ sensors, which is subject to two attacks that control 30% and 20% of the sensors,

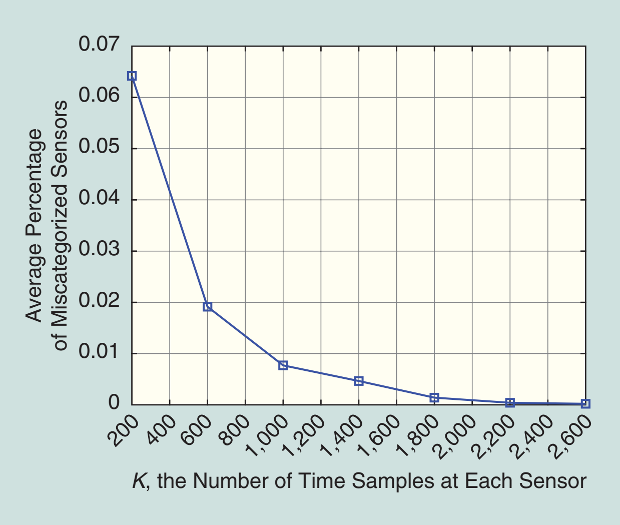


FIGURE 2. Identification and categorization of attacked sensors.

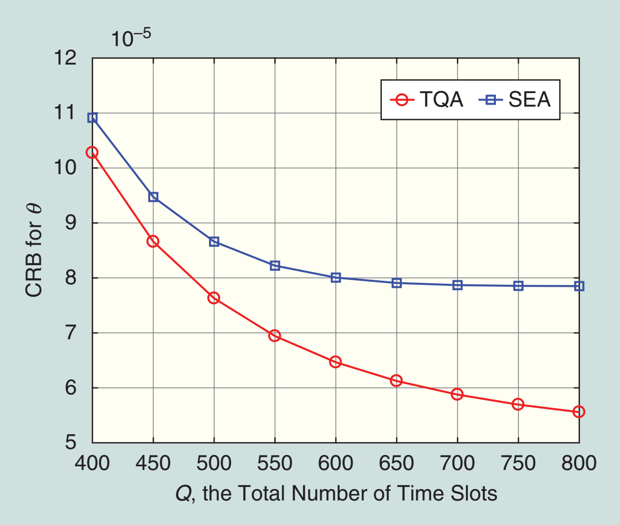


FIGURE 3. The CRB comparison between the TQA and the SEA.

respectively, and modify their observations with attack parameters $(\psi_{1,0}, \psi_{1,1}) = (0.2, 0.8)$ and $(\psi_{2,0}, \psi_{2,1}) = (0.7, 0.1)$. The parameter to be estimated is $\theta = 1$, the threshold of the quantizer in (2) is $\upsilon = 1$, $\Delta_0 = \Delta = 20\%$, and the additive noise obeys a standard normal distribution. In agreement with Theorem 1, Figure 2 depicts a 200-run Monte Carlo approximation of the average percentage of miscategorized sensors that appears to decrease toward zero as the number of time samples at each sensor increases.

Illustrative example: CRB comparison between the TQA and the SEA

Consider a network with $N = 100$ sensors, $\theta = 2$, and two different attacks. The first attack tampers with 25% of the sensors using attack parameters $\psi_{1,0} = 0.9$ and $\psi_{1,1} = 0.95$. The other attack controls 20% of the sensors while using the attack parameters $\psi_{2,0} = 0.15$ and $\psi_{2,1} = 0.2$. The length of each time slot is fixed at $K_t = 10$, and the set of 801 thresholds is $\mathfrak{T} = \{0, -0.125, 0.125, -0.250, 0.250, \dots, -5, 5\}$. All other settings are similar to those of Figure 2. Figure 3 depicts the CRB when estimating θ for the two approaches with varying Q, the total number of time slots. For a given Q, each sensor observes QK_t time samples and picks the first Q thresholds from the set of thresholds $\mathfrak{T}$ to quantize the time samples in different time slots. It is seen that the CRBs for both approaches decrease as Q grows, which is reasonable since the number of time samples at each sensor increases. Moreover, Figure 3 illustrates that the TQA provides significant CRB performance gain when compared to the SEA, which implies that the set of thresholds leads to $\text{rank}(\Xi_p) = 3$ for at least one p based on Theorem 2, and the number of p for which this occurs increases with the increase in Q over the region shown.

Highly desirable spoofing attacks

In the previous section, we essentially described optimum processing of MiMA data for cases with a sufficiently large number of observations. We described how to find which sensors were attacked and how to develop groups of similarly attacked sensors. We also described how and when to use the data at the attacked sensors and when to not use these data. The method we proposed to use the attacked data involved estimating the attack parameters of a model describing the attack. With this model, we can follow accepted estimation theory to develop an estimation procedure using both the unattacked data and the attacked data. We could use, for example, an MLE procedure since we assume a large number of observations. Grouping together the similarly attacked data would help this procedure. We note that one could develop algorithms to automatically do the sensor grouping of similarly attacked sensors, determination of which sensors are unattacked, and MLE using an approach similar to that in [17]. Further, one can extend many of the ideas considered in this section to spoofing attacks; see [12].

In this section, besides considering spoofing attacks, we shift our considerations to find highly desirable attacks from the attacker's point of view. In particular, we are interested in attacks that will guarantee that the after-attack estimation performance must produce a CRB larger than some specified value, regardless of how the estimation system processes the data. To provide insight into spoofing attacks, vector-desired parameter estimation cases, arbitrary nonbinary quantization, and nonidentically distributed samples, we consider all of these in this section.

Let the after-attack unquantized observation $\tilde{x}_{jk}$ be a component of an independent sequence over $(j,k) \in \{1,\dots,N\} \times \{1,\dots,K\}$, and assume each $\tilde{x}_{jk}$ may be exposed to a spoofing attack to yield a pdf that can be expressed as

$$\tilde{x}_{jk} \sim \begin{cases} f_{jk}(\tilde{x}_{jk}|\boldsymbol{\theta}), & \text{if } j \in \mathcal{A}_0, \\ g_{jk}(\tilde{x}_{jk}|\boldsymbol{\theta},\boldsymbol{\tau}^{(p)}), & \text{if } j \in \mathcal{A}_p. \end{cases} \qquad (12)$$

The notations $\tilde{x}_{jk}$ and $\tilde{u}_{jk}$ denote the after-attack unquantized and quantized measurements regardless of whether the jth sensor is attacked or not, respectively. From (12), if $j \in \mathcal{A}_p$ for $p = 1, 2, \dots, P$, then the after-attack pdf $g_{jk}\left(x_{jk}\middle|\boldsymbol{\theta},\boldsymbol{\tau}^{(p)}\right)$ is parametrized by the desired vector parameter $\boldsymbol{\theta}$ with dimension $D_{\boldsymbol{\theta}}$ and the attack vector parameter $\boldsymbol{\tau}^{(p)}$ with dimension D_p.

To conform to previous work on spoofing attacks, the functional forms of the attacks, and equivalently $\{g_{jk}\}$, are assumed known to the attacked system, but the desired and attack vector parameters are not. All existing research considers cases in which the attacker replaces the unattacked sensor data by a function of the unattacked data, where the form of the function is known down to some unknown scalar quantities, which we call *attack parameters*. For example, a specific type of spoofing attack, called a *data-injection attack*, adds an unknown attack parameter to the sensor data. Thus, the function here is a linear function with unit slope, and the attack parameter is the value added to the sensor data.

Along with considering a vector desired parameter, this section generalizes the quantization model to allow nonbinary quantization. At the jth sensor, each after-attack measurement $\tilde{x}_{jk}$ is quantized to $\tilde{u}_{jk}$ by using an R_j-symbol quantizer with quantization regions $\{I_j^{(r)}\}_{r=1}^{R_j}$, that is,

$$\tilde{u}_{jk} = \sum_{r=1}^{R_j} \{\tilde{x}_{jk} \in I_j^{(r)}\} r, \tag{13}$$

where $\{\cdot\}$ is the indicator function. Let

$$\boldsymbol{\Theta} \triangleq [\boldsymbol{\theta}^T, (\boldsymbol{\tau}^{(1)})^T, ..., (\boldsymbol{\tau}^{(P)})^T]^T \tag{14}$$

denote a vector containing the unknown vector parameter $\boldsymbol{\theta}$ along with all of the unknown attack vector parameters that parametrize the spoofing attacks.

Optimal guaranteed degradation spoofing attack

Now we define a highly desirable attack.

Definition 1

Consider attacks imposing $\{f_{jk}(x_{jk}|\boldsymbol{\theta})\}$ and $\left\{g_{jk}\left(\tilde{x}_{jk}|\boldsymbol{\theta}, \boldsymbol{\tau}^{(p)}\right)\right\}$. The optimal guaranteed degradation spoofing attack (OGDSA) maximizes the degradation of the CRB for the vector parameter of interest at the FC when the attacked sensors are well identified and categorized according to distinct types of spoofing attacks by the FC. The CRB for the case where the attacked sensors are well identified and categorized provides a lower bound on the CRB for any case, including cases with unidentified and uncategorized attacked sensors, thus providing guaranteed sufficiently undesirable estimation performance for the estimation system and justifying the name. One class of attacks that are OGDSA are called *inestimable spoofing attacks* (*ISAs*), defined next and further illuminated by Theorem 3.

Definition 2 (Inestimable spoofing attack)

The pth spoofing attack is referred to as an *ISA* if the corresponding FIM for estimating $\boldsymbol{\tau}^{(p)}$ is singular. Such an attack can result from a sufficiently powerful attack relative to the number of quantization symbols employed by the quantizers as quantified by Theorem 3.

Theorem 3

For the pth spoofing attack, if the dimension D_p of the attack parameter $\boldsymbol{\tau}^{(p)}$ satisfies

$$D_p > \sum_{j \in \mathcal{A}_p} K(R_j - 1), \tag{15}$$

then the FIM for estimating $\boldsymbol{\tau}^{(p)}$ is singular, and, furthermore, the FIM for estimating $\boldsymbol{\Theta}$ is also singular.

Recall from the discussion just prior to Theorem 2 that the fixed threshold quantization approach fails for MIMAs due to a singular FIM. Theorem 3 shows that similar failures (certain FIMs become singular) can occur for spoofing attacks. The failures occur because the quantization approach produces data that cannot be used to estimate more parameters than the right-hand side of (15). Thus, if we form an attack that involves more attack parameters than the right-hand side of (15), then the FIM for estimating $\boldsymbol{\tau}^{(p)}$ is singular. To attack the estimation system and cause such a failure, one only needs to map the unattacked data through a function depending on all of the components of the attack parameter vector whose dimension is larger than the right-hand side of (15). A polynomial with coefficients that are the components of the attack parameter vector is one such function. Now, after quantization, an unbiased estimation approach is not capable of estimating the attack parameters to statistically model the attacked data (by modeling the function), so it cannot recover the desired parameter. The other possible class of OGDSAs, called *optimal estimable spoofing attacks* (*OESAs*), are a subset of estimable spoofing attacks (ESAs), defined next.

Definition 3 (ESA)

The pth OGDSA spoofing attack is said to be estimable if the corresponding FIM for estimating $\boldsymbol{\tau}^{(p)}$ is nonsingular. Reference [17] demonstrates that the attacked observations are useless for estimating the desired vector parameter under an OESA. Theorem 4 is useful for catagorizing ESAs.

Theorem 4

In the presence of ESAs, the CRB must satisfy

$$\mathrm{CRB}_{\mathrm{ESA}}(\boldsymbol{\theta}) \triangleq [\mathbf{J}_{\boldsymbol{\Theta}}^{-1}]_{1:D_{\boldsymbol{\theta}}} \preceq \mathbf{J}_{\mathcal{A}_0}^{-1}, \tag{16}$$

where $\mathbf{J}_{\boldsymbol{\Theta}}$ denotes the FIM for estimating $\boldsymbol{\Theta}$, and $[\mathbf{J}_{\boldsymbol{\Theta}}^{-1}]_{1:D_{\boldsymbol{\theta}}}$ is the $D_{\boldsymbol{\theta}}$-by-$D_{\boldsymbol{\theta}}$ leading principal minor of $\mathbf{J}_{\boldsymbol{\Theta}}$. The matrix $\mathbf{J}_{\mathcal{A}_0}$ is the FIM for estimating the desired vector parameter $\boldsymbol{\theta}$ by using only the data from $\mathcal{A}_0$.

In [17], necessary and sufficient conditions are provided for the equality in (16) that ultimately defines the class of OESAs for a given estimation problem. The necessary and sufficient conditions are provided in terms of a relationship between the subspaces spanned by the columns of certain matrices related to the FIMs for estimating $\boldsymbol{\theta}$ and $\boldsymbol{\tau}^{(p)}$ using data under the pth attack. One trivial example of an OESA, which may be relatively easy to detect, is to replace the original measurements at the attacked sensors by some regenerated data obeying a distribution not parametrized by $\boldsymbol{\theta}$. Nontrivial OESAs can also be given. For example, it is also shown in [17] that a generalization of an additive shift in $\boldsymbol{\theta}$, the attack thus replacing $\boldsymbol{\theta}$ by $\boldsymbol{\theta} + \boldsymbol{\tau}^{(p)}$, is always an OESA for any estimation problem. It is clear that such an attack is very hard for the estimation system to deal with since the unattacked estimation algorithm will be capable of estimating $\boldsymbol{\theta} + \boldsymbol{\tau}^{(p)}$, but it cannot resolve $\boldsymbol{\theta}$ and $\boldsymbol{\tau}^{(p)}$ since an uncountable number of choices for $\boldsymbol{\theta}$ and $\boldsymbol{\tau}^{(p)}$ will all lead to

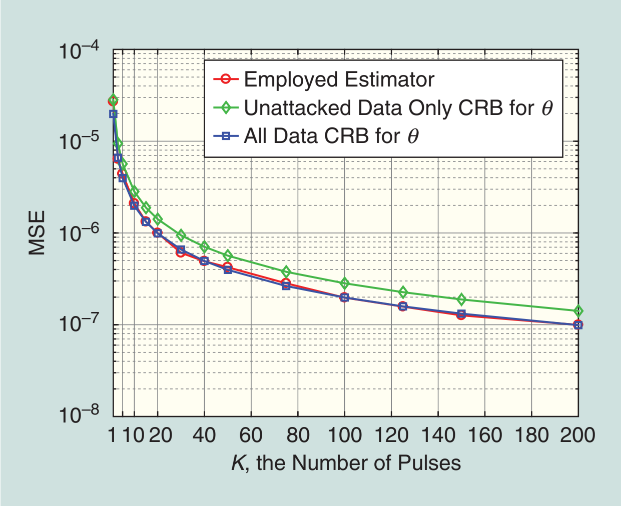


FIGURE 4. The attack performance of the data-injection attacks (non-OGDSA).

the same value of $\boldsymbol{\theta} + \boldsymbol{\tau}^{(p)}$. The estimation system has no way to choose the right one in this settling. On the other hand, other attacks, with different functional forms, can be OESAs for one estimation problem but not for another problem.

It is worth mentioning that, if sensors are correctly categorized, the CRB cannot be worse than the one that ignores the attacked sensors, so the attacked sensors can generally help in terms of reducing CRB. This explains (16) in an intuitive way. However, from the definitions of the ISA and OESA, the OGDSAs essentially make the data from the attacked sensors useless in terms of reducing CRB. Thus they give the equality in (16).

Illustrative example: Comparison between OGDSA and non-OGDSA for multiple-input, multiple-output (MIMO) radar

Previously, we explained how radars are being used in self-driving cars to avoid accidents in which humans and animals might be seriously hurt. Here, we give an example where a radar is spoofed. Consider a multiple-transmitter, multiple-receiver radar (often called a *MIMO radar)* with one transmit station and $N = 10$ receive stations. The first three receive stations are under spoofing attacks. Each station makes M measurements of each pulse in the pulse train and employs an identical 4-bit quantizer with a set of thresholds $\{-\infty, -5, -4, -3, ..., 8, 9, \infty\}$ to convert analog measurements to quantized data before transmitting them to the FC. Without any attack, the mth measurement of the kth pulse in the pulse train at the jth station can be expressed as

$$x_{jm}^{(k)} = \sqrt{E_j}\, a_j s(t_{jm}^{(k)} - \theta_j) + n_{jm}^{(k)}, \tag{17}$$

where θ_j is the desired parameter (delay of the transmitted signal after reflection from the radar target), $m = 1, 2, ..., M$, $k = 1, 2, ..., K$, and K is the total number of pulses in the pulse train. Assume $\{n_{jm}^{(k)}\}$ is an independent and identically distributed zero-mean Gaussian noise sequence with variance $\sigma^2 = 5$. The signal $s(t)$ is a Gaussian pulse signal [31], that is, $s(t) = (2/T^2)^{1/4} \exp(-\pi t^2/T^2)$, and the sampling times are $t_{jm}^{(k)} = (m-1)\Delta t, \forall m = 1, 2, ..., M$. To simplify the model, we assume that the distance between the target and any receiving station is much larger than the distances between every pair of receive stations, and, hence, we can assume that $\theta_j = \theta$ for all j. We set the quantities $T = 0.1$, $\Delta t = 0.001, \theta = 0.02$, and $E_j = 1, a_j = 1$ for all j.

First, we consider the attack performance of a non-OGDSA for this estimation problem, called a *data-injection attack*. If the jth station is under a data-injection attack for $j = 1, 2, 3$, the mth after-attack measurement of the kth pulse in the pulse train is given by

$$\tilde{x}_{jm}^{(k)} = \sqrt{E_j} a_j s(t_{jm}^{(k)} - \theta_j) + \xi_j + n_{jm}^{(k)}, \tag{18}$$

where the attack parameters are $\xi_1 = 1, \xi_2 = -2$, and $\xi_3 = -1$. We employ an expectation-maximum-based joint attack identification and parameter estimation approach proposed in [17] to estimate the desired parameter θ. Figure 4 depicts the MSE performance of the employed estimator plotted on a log scale, where $M = 40$. The clairvoyant CRB for estimating θ, which knows which sensors are attacked and uses data from all sensors, is also plotted in Figure 4 along with the CRB for estimating θ, which uses data only from unattacked sensors. Figure 4 shows that the CRB that uses only the unattacked sensor data is strictly larger than the CRB that uses all the data, which implies that the attacked data are useful for reducing the CRB. As expected, the data-injection attack does not make the attacked data useless for reducing the CRB as opposed to an OGDSA. Moreover, the employed estimation approach can outperform the CRB that uses only the unattacked data and asymptotically achieves the clairvoyant CRB that uses all sensor data.

Next, we consider another spoofing attack, called a *delay attack,* which is a shift-in-parameter OESA (previously discussed and mentioned after Theorem 4). This attack alters the delay in the received signal, possibly by employing DRFM along with a receiver/transmitter [17] to transmit the signal back toward the receive antennas with an arbitrary delay chosen by the attacker. For the jth station, which is under a delay attack for $j = 1, 2, 3$, the mth after-attack measurement of the kth pulse in the pulse train is given by

$$\tilde{x}_{jm}^{(k)} = \sqrt{E_j}\, a_j s(t_{jm}^{(k)} - \theta_j - \xi_j) + n_{jm}^{(k)}, \tag{19}$$

where ξ_j is the delay introduced by the delay attack. It can be shown that the delay attack in (19) is an OGDSA [17], which is also an OESA. In Figure 5, the simulation setting is the same as that in Figure 4, except $M = 3$ and the attack parameters are $\xi_1 = 0.04, \xi_2 = 0.05$, and $\xi_3 = 0.06$. We employ the same estimation approach as that employed in Figure 4. Figure 5 illustrates the MSE performance of the employed estimator along with the CRB for θ, which knows which sensors are attacked and uses data only from unattacked sensors. It is worth mentioning that the employed approach can perfectly identify the attacked sensors with large K [17], and it is seen that the MSE performance of the employed estimator converges to the CRB using only unattacked data. Most importantly, the large K results in Figure 5 agree with the previously stated theoretical results saying the attacked data are not useful in reducing the CRB under an OGDSA.

General attacks in vector parameter estimation systems

In the sections "MiMAs" and "Highly Desirable Spoofing Attacks," we considered MiMAs and spoofing attacks separately. In this section, we consider the most general attacks, which include combinations of MiMAs and spoofing attacks, when estimating the location of an acoustic emitter [6] at $\boldsymbol{\zeta}_{\mathrm{T}} = [y_{\mathrm{T}}, z_{\mathrm{T}}]$, where y_{T} and z_{T} denote the coordinates of the emitter location in the two-dimensional plane. We assume that the emitter is in some region of interest (ROI) $\mathcal{S}$. For the jth sensor, we use $\boldsymbol{\zeta}_j = [y_j, z_j]$ to denote its location. In addition to N insecure sensors, the estimation system has access to two secure sensors, considered the $(N+1)$th and $(N+2)$th sensors, respectively. These two secure sensors are well protected and thereby are guaranteed to be unattacked, while the other N sensors are open to attacks. We assume that the signal radiated from the emitter obeys an isotropic power attenuation model [6] and each sensor observes K data samples. The kth data sample at the jth sensor is described as $x_{jk} = P_0(D_0/D_j)^{\gamma} + n_{jk},\ j = 1,2,...,N+2,$ where the distance D_j between the jth sensor and the emitter is defined by $D_j \overset{\Delta}{=} \| \boldsymbol{\zeta}_j - \boldsymbol{\zeta}_{\mathrm{T}} \| = \sqrt{(y_j - y_{\mathrm{T}})^2 + (z_j - z_{\mathrm{T}})^2}, \forall j$, the quantity P_0 is the power measured at a reference distance D_0, γ is the path-loss exponent that is a positive constant, and n_{jk} denotes the additive noise sample with pdf $f_j(n_{jk})$. We assume that P_0, D_0, γ, $\{f_j(\cdot)\}_{j=1}^{N+2}$, and $\{\boldsymbol{\zeta}_j\}_{j=1}^{N+2}$ are known to the FC. Moreover, we assume $\{n_{jk}\}$ are independent and, for each j, $\{n_{jk}\}_{k=1}^{K}$ is an identically distributed sequence.

Each sensor j quantizes its sample x_{jk} to one-bit data u_{jk} by using the threshold v_j, and then transmits u_{jk} to the FC, that is, $u_{jk} \overset{\Delta}{=} \ \{x_{jk} \in (\nu_j, \infty)\}, \forall j$ and $\forall k$, where $\ \{\cdot\}$ is the indicator function. We assume that the thresholds $\{\nu_j\}_{j=1}^{N+2}$ are known to the FC.

If $j \in \mathcal{A}_p$ for some $p \geq 1$, the after-attack quantized data can be generally expressed as $\tilde{u}_{jk} = \tilde{h}_{jk}(\ \{h_{jk}(x_{jk}) \in (\nu_j, \infty)\})$, where the maps $h_{jk}(\cdot)$ and $\tilde{h}_{jk}(\cdot)$ represent the effects of the spoofing attack and the MiMA at time k, respectively. Similar to (2), the NMLE of the distance D_j can be expressed as

$$\hat{D}_j^{(K)} = D_0 P_0^{\frac{1}{\gamma}} \left[\nu_j - F_j^{-1} \left(\frac{1}{K} \sum_{k=1}^{K} (1 - \tilde{u}_{jk}) \right) \right]^{-\frac{1}{\gamma}}, \qquad (20)$$

which yields that, for the two secure sensors, that is, the $(N+1)$th and $(N+2)$th sensors, we have

$$\hat{D}_{N+1}^{(K)} \to D_{N+1} \text{ and } \hat{D}_{N+2}^{(K)} \to D_{N+2} \text{ almost surely, as } K \to \infty, \qquad (21)$$

since $\tilde{u}_{jk} = u_{jk}$ for $j = N+1$ and $N+2$. Based on this fact, we can generate two circles that are centered at the $(N+1)$th and $(N+2)$th sensors with radii equal to $\hat{D}_{N+1}^{(K)}$ and $\hat{D}_{N+2}^{(K)}$, respectively. In the asymptotic regime, where $K \to \infty$, the intersection point of these two circles pinpoints the location of the emitter under the assumption that the ROI $\mathcal{S}$ is contained in one of the two half spaces produced by dividing the whole space by the line passing through the two secure sensors. Similarly, if the jth sensor is unattacked (attacked), the circle centered at the jth sensor with radius equal to $\hat{D}_j^{(K)}$ should (should not) pass through this intersection point in the asymptotic regime where $K \to \infty$. Thus, we can determine whether the jth sensor is attacked or not by checking this geometric consistency among the circles associated with the two secure sensors and the jth sensor in the asymptotic regime where $K \to \infty$.

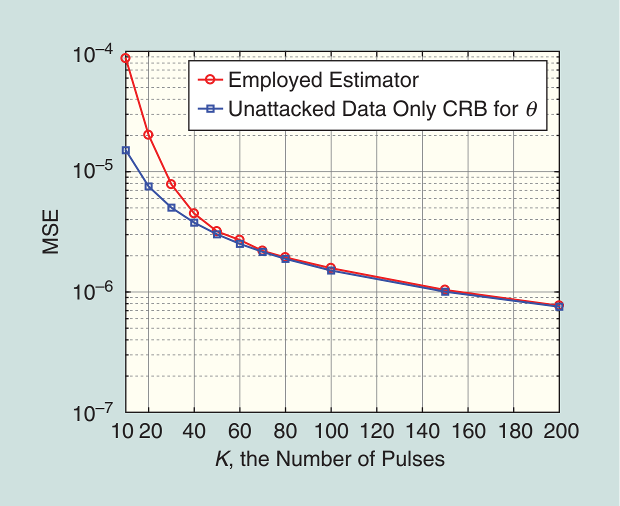


FIGURE 5. The attack performance of the DRFM attacks (OGDSA).

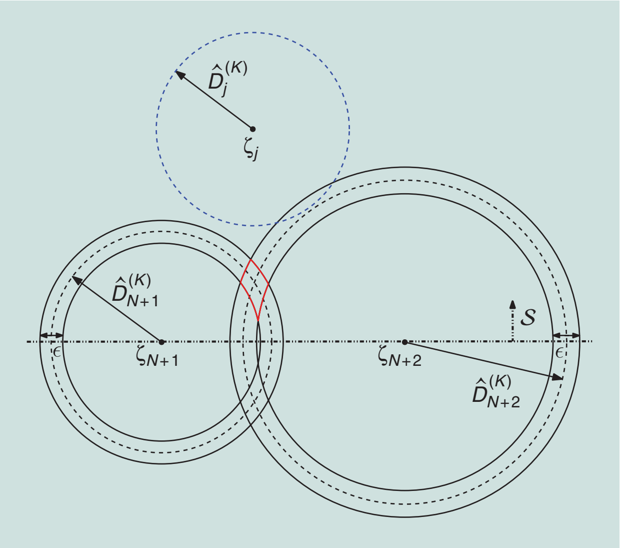


FIGURE 6. A geometric illustration of the proposed detectors.

In the regime where K is finite, the attack-detection procedure is similar except that, for each of the two secure sensors, the associated circle is replaced by a ring with some constant width ε; see Figure 6. We declare that the jth sensor is unattacked (attacked) if the circle (the blue dashed circle in Figure 6) associated with the jth sensor passes (does not pass) through the overlap area (the area enclosed by the red curves in Figure 6) of the rings associated with the two secure sensors. In the exact situation in

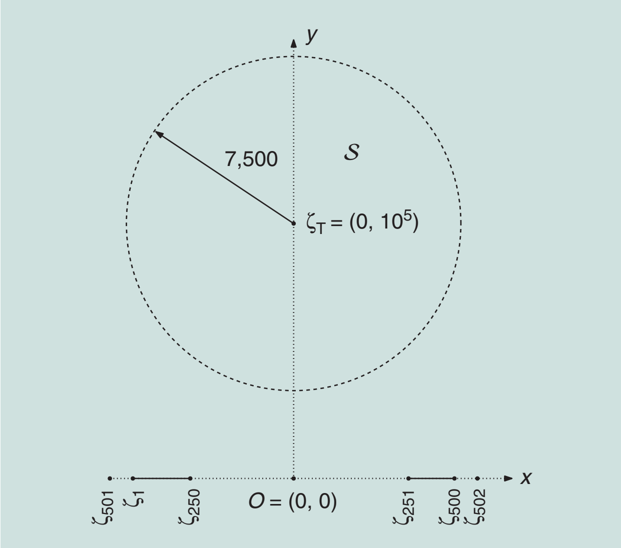


FIGURE 7. The simulation configuration.

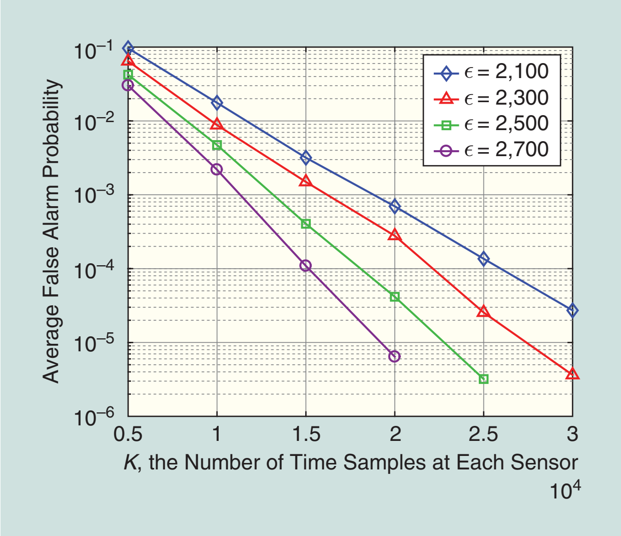


FIGURE 8. The false alarm probability for different ϵ.

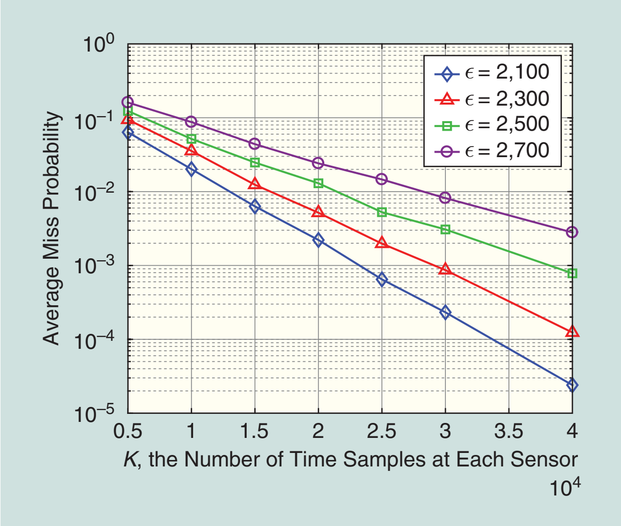


FIGURE 9. The miss probability for different ϵ.

Figure 6, we would declare an attack. The mathematical formulation of this attack-detection idea can be found in [12].

By employing large deviations principles, we derive the following theorem regarding the performance of the proposed detectors.

Theorem 5

If widths of the rings associated with the two secure sensors are smaller than C_0, where C_0 is a constant defined in [12], then the false alarm and miss probabilities are upper bounded by two exponentially decaying functions of K, respectively. The rates of decay can also be found in [12].

The idea of detecting attacks in this emitter localization problem can be generalized to the general IoT sensor network estimation problem equipped with secure sensors. In particular, we can employ the data from the secure sensors to generate some constraints that are satisfied by the desired parameters with high probability (that the desired parameters must lie in the overlap of two rings for the just-described localization example was one such constraint). Then we can detect whether or not each insecure sensor is attacked by checking whether or not the NMLE based on the data from the sensor satisfies the constraints.

Illustrative example: Proposed detector for general attacks

To illustrate Theorem 5, we test the performance of the proposed detector for an example case. The system configuration is illustrated in Figure 7. Consider a network consisting of two groups of sensors with $N = 500$. The two secure sensors are located at $\boldsymbol{\zeta}_{501} = (-10^3, 0)$ and $\boldsymbol{\zeta}_{502} = (10^3, 0)$, respectively. The rest of the sensors are all located along the x-axis and are partitioned into two groups. In the first group, the sensors $\{1, 2, ..., 250, 501\}$ are evenly spaced between $(-10^3, 0)$ and $(-0.9 \times 10^3, 0)$, while sensors in second group $\{251, 252, ..., 500, 502\}$ are evenly spaced between $(0.9 \times 10^3, 0)$ and $(10^3, 0)$. The ROI $\mathcal{S}$ is a disc centered at $(0, 10^5)$ and with radius equal to 7,500. The emitter is located at $\boldsymbol{\zeta}_T = (0, 10^5)$. We assume that $P_0 = 1, D_0 = 10^5$, and $\gamma = 2$. The thresholds $\nu_j = 1$ for all j and n_{jk} follow a Gaussian distribution with zero mean and unit variance. We assume that 250 sensors $\{1, 2, ..., 250\}$ are under a MiMA as described in (3) with $\psi_{j,0} = 0$ and $\psi_{j,1} = 0.94$ for $j = 1, 2, ..., 250$. The rest of the sensors are unattacked.

The average false alarm and miss probabilities versus K are depicted on a log scale in Figures 8 and 9 for four detectors with $\epsilon = 2,100; 2,300; 2,500;$ and $2,700,$ respectively. Figures 8 and 9 show that, for each detector, the average false alarm and miss probabilities decrease exponentially as K grows, which agrees with the theoretical results in Theorem 5. Moreover, as illustrated in Figure 8, the larger the value of ϵ, the smaller the average false alarm probability. On the other hand, Figure 9 shows that the larger the value of ϵ, the larger the average miss probability. Thus, the proper tradeoff between the false alarm and miss probabilities can be chosen by adjusting the value of ϵ.

Implications for unattacked systems

Motivated by Theorem 3, a fundamental limitation on quantized estimation systems not under attack is uncovered. Before proceeding, we first provide two definitions on the identifiability of a vector parameter point and a vector parameter space. Let $\mathbf{\Omega} \subseteq \mathbb{R}^{D_\theta}$ denote the parameter space of interest with a nonempty interior.

Definition 4 (Identifiable vector parameter point)

The vector parameter point $\boldsymbol{\theta} \in \mathbf{\Omega}$ is called identifiable if the conditional distribution of the data conditioned on $\boldsymbol{\theta}$ is not identical to that for any other vector parameter point in $\mathbf{\Omega}$.

Definition 5 (Identifiable vector parameter space)

The vector parameter space $\mathbf{\Omega}$ is considered identifiable if every vector parameter point in $\mathbf{\Omega}$ is identifiable.

Under some mild assumptions [32], we can derive Theorem 6 on the fundamental limitation of quantized estimation systems.

Theorem 6

Let $D_{\boldsymbol{\theta}}$ be the dimension of a vector parameter in $\mathbf{\Omega}$ we want to estimate from L independent observations quantized using Q distinct quantizer designs with R_j, $j = 1,2,\ldots,Q$ symbols. Assume the *j*th group of observations, all quantized by an identical quantizer, are generated from M_j different pdfs. If

$$D_{\boldsymbol{\theta}} > \sum_{j=1}^{Q} M_j(R_j - 1), \tag{22}$$

then the FIM is singular, and, moreover, the vector parameter space $\mathbf{\Omega}$ is not identifiable.

In addition, for any open subset $O \subseteq \mathbf{\Omega}$ in $\mathbb{R}^{D_\theta}$, there are infinitely many vector parameter points in O that are not identifiable.

For identical $(Q = 1)$ binary $(R_j = 1)$ quantization at each sensor and identically distributed observations $(M_1 = 1)$ at each sensor, $\Sigma_{j=1}^{Q} M_j(R_j - 1) = 1$, so a scalar parameter $(D_\theta = 1)$ alone will not satisfy the sufficient condition in (22) for FIM singularity in this case. Note that we have already given just such an example in the section "MiMAs" $(\Sigma_{j=1}^{Q} M_j(R_j - 1) = 1)$ in the discussion just prior to Theorem 2, where the fixed threshold quantization approach worked well (no singular FIM) when there was no attack, since we were estimating a scalar parameter θ. However, the approach failed (singular FIM) with the attack since the parameter to estimate had dimension three and, yet, the right-hand side of (22) is exactly one.

Note that the quantity $\Sigma_{j=1}^{Q} M_j(R_j - 1)$ does not depend on the quantization regions, the number L of observations, the pdfs that generate the observations, or the particular estimation problem but instead depends only on the number of different pdfs involved, the number of quantizers employed, and the number of quantization symbols. This critical quantity is referred to as the *IDQD*.

Theorem 6 reveals a fundamental limitation when utilizing quantized data for estimating a vector parameter and sheds light on the preliminary design of a quantized estimation system. To be specific, the quantization and sensing approach employed should guarantee that the IDQD of the quantized estimation system is larger than or equal to the dimension of the vector parameter of interest. For some specific estimation problems, the singularity of the FIM and the nonidentifiability of the parameter space can exist even if the condition in (22) does not hold.

In some cases, where $D_{\boldsymbol{\theta}}$ is larger than the IDQD, all vector parameter points in $\mathbf{\Omega}$ are nonidentifiable, while in some other cases, there exist some parameter points in $\mathbf{\Omega}$ that are identifiable. Thus, a singular FIM does not necessarily determine the nonidentifiability of the parameter point though it does determine the nonidentifiability of the parameter space. Moreover, it can be shown that, if $D_{\boldsymbol{\theta}}$ is larger than the IDQD, the cardinality of a set of parameter points such that the conditional distribution of the data conditioned on the parameter is identical to that for some other parameter point can be as small as one and can also be uncountably infinite. Generalized results that do not require some assumptions in Theorem 6, e.g., independence, can be found in [32].

Some recent work on related protection layers employing signal processing

Many IoT applications, for example, smart grid and manufacturing, require sensor data to be sent from one location to a different location so the data can be used to change a control or reconfigure the grid or manufacturing process. To provide the low latency required to avoid unstable control loops, low-complexity encryption approaches have received attention for estimation using sensor data in the IoT. An interesting low-complexity approach to encrypt binary quantized data was suggested in [33], which is called *stochastic encryption*. The basic idea is to flip the binary data using an approach similar to our attack model in (4). Then the desired user, who knows the flipping probabilities, will use a maximum likelihood decoding approach to estimate the desired parameter, θ, in (1). The estimation performance loss due to not knowing how each bit was flipped but knowing only the flipping probabilities is shown to be small in [33] with proper design.

It is also shown in [33] that any eavesdroppers will have very poor estimation performance for properly chosen flipping probabilities. The flipping probabilities act as an encryption key for a very low-complexity encryption process that is suitable for a low-complexity sensor node. Using Theorem 6, we have shown [34] that the approach in [33] can only estimate a scalar parameter, so in [34] and [35], we generalized the approach by using different quantizers and flipping probabilities at each sensor, which can also employ nonbinary quantization. Based on Theorem 6, such approaches can be designed to potentially estimate vector parameters of any size while retaining the advantages of the approach suggested in [33]. Now, one might think that, after observing a sufficiently large window of data, an eavesdropper might be able to estimate the flipping probabilities and break the code to estimate the parameter of interest. In [34], we show this is not possible if the eavesdropper employed an unbiased estimator based on Theorem 6. The quantization approach is not of sufficient complexity to allow the eavesdropper to estimate all the quantities needed for him or her to develop an accurate estimate of the parameter of interest.

Stochastic encryption was also considered for defending against eavesdroppers in the context of sequential hypothesis testing in [36]. Since the flipping probabilities are known only to the desired user but not to the eavesdropper, the desired user employs the optimal sequential probability ratio test (SPRT) for sequential detection, whereas the eavesdropper employs a mismatched SPRT. However, every stochastic encryption degrades the performance of the SPRT at the desired user by increasing the expected sample size. In [36], an optimal stochastic encryption is obtained analytically in the sense of maximizing the difference between the expected sample sizes required at the eavesdropper and the desired user, provided that the acceptable tolerance of the increase in the expected sample size at the desired user induced by the stochastic encryption is small enough.

We next describe a technology based on information theory that can provide additional layers of protection that has received recent attention. All communications networks are designed using layers that are different than the security layers we mentioned previously. The lowest layer of a communication network is the physical layer. Most currently employed security procedures are implemented in the network or higher layers, which are a few layers above the physical layer. Since one can do things at the physical layer that cannot be done at the higher layers, the idea of physical layer security seems very attractive. Using physical layer security for the right situations, one can design signals that ensure a required information rate is received by the desired user, but the rates received by the eavesdropper are guaranteed to be less than some small value. Such results exploit information-theoretic ideas and have been called *information-theoretic secrecy.*

The seminal work of Shannon [37] and Wyner [38] laid the foundation for physical layer security, by providing basic formalisms for security in cipher systems and wiretap channels, respectively. Csiszár and Körner generalized Wyner's work to the broadcast channel with confidential messages in [39], which provides a model that aids in the understanding of security in wireless systems. Excellent surveys on physical layer security can be found in [40]–[42]. Authentication, a counterpart of secrecy, has also been given a physical layer security treatment. The study of authentication in an information-theoretic context began with [43] and was extended to the physical layer by Lai et al. in [44]. Besides the information-theoretic investigations, much work has also been done with authentication at the physical layer with practical schemes that utilize the characteristics of the channel and the communication devices to uniquely identify sources. A survey on this topic can be found in [45], and practical methods for wireless authentication utilizing fingerprint embedding at the physical layer can be found in [46] and [47].

Conclusions

In this article, the estimation of an unknown deterministic scalar parameter in the presence of MiMAs has been introduced first. The capability of the IoT systems, in terms of identifying and categorizing the attacked sensors into different groups according to distinct types of attacks, has been outlined in the face of MiMAs. Necessary and sufficient conditions have been provided under which utilizing the attacked sensor data will lead to a more favorable CRB when compared to approaches where the attacked sensors are ignored. Next, necessary and sufficient conditions have been provided under which spoofing attacks provide a guaranteed attack performance in terms of the CRB for estimating a deterministic parameter vector regardless of the processing the estimation system employs. It has been shown that it is always possible to construct such a highly desirable attack by properly employing an attack vector parameter having a sufficiently large dimension relative to the number of quantization symbols employed, which had not been observed previously. In addition, the most general attacks, which include combinations of MiMAs and spoofing attacks, have been considered in an emitter localization system. Attack detectors have been proposed whose false alarm and miss probabilities decrease exponentially as the number of measurement samples increases. For unattacked quantized estimation systems, a general limitation on the dimension of a vector parameter that can be accurately estimated has been uncovered. References that provide various extensions to all of the specific results presented in this article have been supplied, and a brief discussion of low-complexity encryption and physical layer security has been provided.

Acknowledgments

We would like to thank Jake Perazzone for his helpful discussions on physical layer security. This material is based upon work partially supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under grant number W911NF-17-1-0331 and by the National Science Foundation under grants ECCS-1647198, ECCS-1744129, CNS-1702555, and CNS-1702808.

Authors

Jiangfan Zhang (jz2833@columbia.edu) received the B.Eng. degree in communication engineering from Huazhong University of Science and Technology, Wuhan, China, in 2008, the M.Eng. degree in information and communication engineering from Zhejiang University, Hangzhou, China, 2011, and the Ph.D. degree in electrical engineering from Lehigh University, Bethlehem, Pennsylvania, in 2016. He is currently a postdoctoral research scientist in the Department of Electrical Engineering at Columbia University, New York. He is a recipient of the Dean's Doctoral Student Assistantship, Gotshall Fellowship, and a P.C. Rossin Doctoral Fellow at Lehigh University.

Rick S. Blum (rblum@lehigh.edu) is the Robert W. Wieseman Professor of Electrical Engineering at Lehigh University, Bethlehem, Pennsylvania. His research interests include signal processing/machine learning for cybersecurity, smart grid, communications, sensor networking, radar, and sensor processing. He is a Fellow of the IEEE, an IEEE Signal Processing Society Distinguished Lecturer, an IEEE Third Millennium Medal winner, and a member of Eta Kappa Nu and Sigma Xi, and he holds several patents. He was awarded an Office of Naval Research Young Investigator Award and a National Science Foundation Research Initiation Award. He was an associate editor of several IEEE transactions and special issues.

H. Vincent Poor (poor@princeton.edu) is the Michael Henry Strater University Professor of Electrical Engineering

at Princeton University, New Jersey. His interests include information theory and signal processing, with applications in wireless networks, energy systems, and related fields. He is an IEEE Fellow, a member of the U.S. National Academy of Engineering and the U.S. National Academy of Sciences, and a foreign member of the Chinese Academy of Sciences and the Royal Society. He received the Technical Achievement and Society Awards of the IEEE Signal Processing Society in 2007 and 2011, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal and a D.Sc. *honoris causa* from Syracuse University, also in 2017.

References

[1] A. Mukherjee, "Physical-layer security in the Internet of Things: Sensing and communication confidentiality under resource constraints," *Proc. IEEE*, vol. 103, no. 10, pp. 1747–1761, 2015.

[2] H. V. Poor and R. F. Schaefer, "Wireless physical layer security," *Proc. Nat. Acad. Sci.*, vol. 114, no. 1, pp. 19–26, 2017.

[3] S. Marano, V. Matta, and L. Tong, "Distributed detection in the presence of Byzantine attacks," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 16–29, 2009.

[4] A. S. Rawat, P. Anand, H. Chen, and P. K. Varshney, "Collaborative spectrum sensing in the presence of Byzantine attacks in cognitive radio networks," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 774–786, 2011.

[5] X. He, H. Dai, and P. Ning, "A Byzantine attack defender in cognitive radio networks: The conditional frequency check," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2512–2523, 2013.

[6] A. Vempaty, O. Ozdemir, K. Agrawal, H. Chen, and P. Varshney, "Localization in wireless sensor networks: Byzantines and mitigation techniques," *IEEE Trans. Signal Processing*, vol. 61, no. 6, pp. 1495–1508, Mar. 2013.

[7] A. Vempaty, L. Tong, and P. Varshney, "Distributed inference with Byzantine data: State-of-the-art review on data falsification attacks," *IEEE Signal Process. Mag.*, vol. 30, no. 5, pp. 65–75, 2013.

[8] B. Kailkhura, S. Brahma, Y. S. Han, and P. K. Varshney, "Distributed detection in tree topologies with Byzantines," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3208–3219, 2014.

[9] B. Kailkhura, V. S. S. Nadendla, and P. K. Varshney, "Distributed inference in the presence of eavesdroppers: A survey," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 40–46, 2015.

[10] J. Zhang, R. S. Blum, X. Lu, and D. Conus, "Asymptotically optimum distributed estimation in the presence of attacks," *IEEE Trans. Signal Process.*, vol. 63, no. 5, pp. 1086–1101, Mar. 2015.

[11] B. Alnajjab, J. Zhang, and R. S. Blum, "Attacks on sensor network parameter estimation with quantization: Performance and asymptotically optimum processing," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6659–6672, Dec. 2015.

[12] J. Zhang, X. Wang, R. S. Blum, and L. M. Kaplan, "Attack detection in sensor network target localization systems with quantized data," *IEEE Trans. Signal Process.*, vol. 66, no. 8, pp. 2070–2085, Apr. 2018.

[13] T. T. Kim and H. V. Poor, "Strategic protection against data injection attacks on power grids," *IEEE Trans. Smart Grid*, vol. 2, no. 2, pp. 326–333, 2011.

[14] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *IEEE Trans. Smart Grid*, vol. 2, no. 4, pp. 645–658, 2011.

[15] S. Cui, Z. Han, S. Kar, T. T. Kim, H. V. Poor, and A. Tajer, "Coordinated data-injection attack and detection in the smart grid: A detailed look at enriching detection solutions," *IEEE Signal Process. Mag.*, vol. 29, no. 5, pp. 106–115, 2012.

[16] D. He, S. Chan, and M. Guizani, "Cyber security analysis and protection of wireless sensor networks for smart grid monitoring," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 98–103, Dec. 2017.

[17] J. Zhang, R. S. Blum, L. M. Kaplan, and X. Lu, "Functional forms of optimum spoofing attacks for vector parameter estimation in quantized sensor networks," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 705–720, 2017.

[18] Z. Li, W. Trappe, Y. Zhang, and B. Nath, "Robust statistical methods for securing wireless localization in sensor networks," in *Proc. 4th Int. Symp. Information Processing Sensor Networks,* Apr. 2005, pp. 91–98.

[19] J. H. Lee and R. Buehrer, "Characterization and detection of location spoofing attacks," *J. Commun. Netw.*, vol. 14, no. 4, pp. 396–409, Aug. 2012.

[20] P. Pradhan, K. Nagananda, P. Venkitasubramaniam, S. Kishore, and R. S. Blum, "GPS spoofing attack characterization and detection in smart grids," in *Proc. IEEE Conf. Communications and Network Security,* 2016, pp. 391–395.

[21] C. Wilson and V. Veeravalli, "MMSE estimation in a sensor network in the presence of an adversary," in *Proc. IEEE Int. Symp. Information Theory,* 2016, pp. 2479–2483.

[22] J. C. Balda, A. Mantooth, R. Blum, and P. Tenti, "Cybersecurity and power electronics: Addressing the security vulnerabilities of the internet of things," *IEEE Power Electron. Mag.*, vol. 4, no. 4, pp. 37–43, Dec. 2017.

[23] V. Nadendla, Y. S. Han, and P. K. Varshney, "Distributed inference with M-ary quantized data in the presence of Byzantine attacks," *IEEE Trans. Signal Process.*, vol. 62, no. 10, pp. 2681–2695, 2014.

[24] Y. Zhao, A. Goldsmith, and H. V. Poor, "Minimum sparsity of unobservable power network attacks," *IEEE Trans. Autom. Control*, vol. 62, no. 7, pp. 3354–3368, 2017.

[25] A. Vempaty, K. Agrawal, H. Chen, and P. Varshney, "Adaptive learning of Byzantines' behavior in cooperative spectrum sensing," in *Proc. IEEE Wireless Communications and Networking Conf.,* 2011, pp. 1310–1315.

[26] D. Liu, P. Ning, A. Liu, C. Wang, and W. K. Du, "Attack-resistant location estimation in wireless sensor networks," *ACM Trans. Inf. Syst. Security*, vol. 11, no. 4, p. 22, 2008.

[27] E. Bland. (2008). GPS 'spoofing' could threaten national security. [Online]. Available: http://www.nbcnews.com/id/26992456

[28] A. Couts. (2013). Want to see this $80 million super yacht sink? With GPS spoofing, now you can! [Online]. Available: http://www.digitaltrends.com/mobile/gps-spoofing/

[29] M. I. Skolnik, *Introduction to Radar Systems*, 2nd ed. New York: McGraw Hill Book Co., 1980.

[30] S. Roome, "Digital radio frequency memory," *Electron. Commun. Eng. J.*, vol. 2, no. 4, pp. 147–153, Aug. 1990.

[31] Q. He, R. S. Blum, and A. M. Haimovich, "Noncoherent MIMO radar for location and velocity estimation: More antennas means better performance," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3661–3680, 2010.

[32] J. Zhang, R. S. Blum, L. Kaplan, and X. Lu, "A fundamental limitation on maximum parameter dimension for accurate estimation with quantized data," arXiv Preprint, arXiv:1605.07679, 2016.

[33] T. C. Aysal and K. E. Barner, "Sensor data cryptography in wireless sensor networks," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 2, pp. 273–289, 2008.

[34] A. N. Samudrala and R. S. Blum, "On the estimation and secrecy capabilities of stochastic encryption for parameter estimation in IOT," in *Proc. IEEE Annu. Conf. Information Science and Systems*, 2018, pp. 1–6.

[35] A. N. Samudrala and R. S. Blum, "Asymptotic analysis of a new low complexity encryption approach for the Internet of Things, smart cities and smart grid," in *Proc. IEEE Int. Conf. Smart Grid and Smart Cities*, 2017, pp. 200–204.

[36] J. Zhang and X. Wang, "Asymptotically optimal stochastic encryption for quantized sequential detection in the presence of eavesdroppers," arXiv Preprint, arXiv:1703.02141, 2017.

[37] C. E. Shannon, "Communication theory of secrecy systems," *Bell Labs Tech. J.*, vol. 28, no. 4, pp. 656–715, 1949.

[38] A. D. Wyner, "The wire-tap channel," *Bell Labs Tech. J.*, vol. 54, no. 8, pp. 1355–1387, 1975.

[39] I. Csiszár and J. Korner, "Broadcast channels with confidential messages," *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 339–348, 1978.

[40] H. V. Poor and R. F. Schaefer, "Wireless physical layer security," *Proc. Nat. Acad. Sci.*, vol. 114, no. 1, pp. 19–26, 2017.

[41] M. Bloch, J. Barros, M. R. Rodrigues, and S. W. McLaughlin, "Wireless information-theoretic security," *IEEE Trans. Inf. Theory*, vol. 54, no. 6, pp. 2515–2534, 2008.

[42] M. Bloch and J. Barros, *Physical-Layer Security: From Information Theory to Security Engineering.* Cambridge, U.K.: Cambridge Univ. Press, 2011.

[43] G. J. Simmons, "Authentication theory/coding theory," in *Proc. Workshop Theory and Application Cryptographic Techniques*, 1984, pp. 411–431.

[44] L. Lai, H. El Gamal, and H. V. Poor, "Authentication over noisy channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 906–916, 2009.

[45] K. Zeng, K. Govindan, and P. Mohapatra, "Non-cryptographic authentication and identification in wireless networks [security and privacy in emerging wireless networks]," *IEEE Wireless Commun.*, vol. 17, pp. 56–62, Oct. 2010.

[46] J. B. Perazzone, P. L. Yu, B. M. Sadler and R. S. Blum, "Cryptographic side-channel signaling and authentication via fingerprint embedding," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2216–2225, Sept. 2018.

[47] H. Liu, Y. Wang, J. Liu, J. Yang, Y. Chen and H. V. Poor, "Authenticating users using fine-grained channel information," *IEEE Trans. Mobile Comput.*, vol. 17, no. 2, pp. 251–264, 2018.

Yuan Chen, Soummya Kar, and José M.F. Moura

The Internet of Things

Secure distributed inference

INTERNET OF THINGS—ISTOCKPHOTO.COM/IAREMENKO
CIRCUITS—IMAGE LICENSED BY INGRAM PUBLISHING

The growth in the number of devices connected to the Internet of Things (IoT) poses major challenges in security. The integrity and trustworthiness of data and data analytics are increasingly important concerns in IoT applications. These are compounded by the highly distributed nature of IoT devices, making it infeasible to prevent attacks and intrusions on all data sources. Adversaries may hijack devices and compromise their data. As a result, reactive countermeasures, such as intrusion detection and resilient analytics, become vital components of security. This article overviews algorithms for secure distributed inference in IoT.

Introduction

As the number of devices connected to the IoT continues to grow, the security of data generated, processed, and transceived by these devices becomes a pressing issue. IoT applications feature connected heterogeneous devices that share a common overarching goal; they cooperate and exchange information to complete their objective. For example, in a smart home, a car may communicate with the garage to automatically open the door, and wearable gadgets may exchange information with smart thermostats and lighting fixtures to create a comfortable environment [1]. Connected automobiles in vehicular networks use onboard sensor measurements to monitor road conditions, find open parking spaces, and estimate traffic patterns [2], [3]. In mobile crowdsensing, individuals use their smartphones to monitor noise levels in neighborhoods and estimate wait times for public transportation [4]. In the smart grid, smart electricity meters make real-time measurements of power demand and consumption and are vital components of optimal power dispatch [5]. Smart cities are instrumented with sensors to observe traffic, monitor weather, and measure air quality [1], [6], [7].

Certain IoT applications feature devices specifically for monitoring and controlling physical systems. Machines on an assembly line may be equipped with sensors to detect production anomalies and predict when parts need to be replaced [1]. Phasor measurement units, smart electricity meters, circuit

Digital Object Identifier 10.1109/MSP.2018.2842097
Date of publication: 28 August 2018

1053-5888/18©2018IEEE

breakers, and generation sources monitor and control the state of the smart grid [5], [8]. Other applications simply feature a collection of devices that need to cooperate to complete a shared task. For example, in a smart home, a smart speaker (e.g., Google Home, Amazon Echo) is responsible for deciphering a user's voice commands and relaying this information to a television for broadcasting a movie, to lighting fixtures for dimming the lights, or to the thermostat for changing the temperature. The common characteristic of all of these IoT applications is that the devices must cooperatively process information to accomplish their collective objective, whether it is monitoring a physical system or dimming light features using voice commands.

A key task in these applications is inference, processing measurement data for information. The quality of inference critically depends on the integrity of the sensor data, i.e., on the trustworthiness of the sensors and devices producing the data. IoT devices, ranging in scale from pacemakers, to cars, to phasor measurement units (PMUs) in the smart grid are vulnerable to cyberattacks [1], [8], [9]. Malicious adversaries may hijack devices, arbitrarily corrupt their data, jam communication links, and mislead the application to produce erroneous inferences.

In this article, we overview secure inference for the IoT and highlight cooperation strategies that are resilient to data integrity attacks. Previous work has surveyed data analysis in the IoT without adversaries [10], reviewed protocols for secure data communication [11], and summarized security challenges in the IoT [12]. Reference [8] surveys recent advances in security for the smart grid and presents a broad summary of secure data acquisition, communication, storage, and processing. In contrast, this article provides a focused, more detailed overview of secure data processing and inference for the IoT.

We consider three main architectures of IoT systems. In centralized architectures, a single entity processes all of the data from all of the devices. In decentralized or parallel architectures, devices perform local processing and transmit the processed data to a fusion center, which completes the computation task [13]. In fully distributed architectures, individual devices cooperate with neighbors over a communication network and perform all of the processing. We present algorithms for each of these architectures that fuse data streams from many separate devices and still produce a collective accurate inference even while an adversary tampers with a subset of the devices and manipulates their data streams.

The end goal is to correctly process the data even in the presence of adversaries. One strategy to achieve this goal is by detecting and identifying attacks and, after doing so, taking corrective action. An alternative approach is to use resilient processing algorithms, which, by design, resist attacks without explicit detection and identification. This article overviews both types of strategies. The techniques we present are generic and not application specific; for illustrative purposes, however, we will explain these techniques in the context of air-quality monitoring (Figure 1). In practice, one is interested in graceful degradation of performance in the presence of adversaries. The results and approaches we consider here, due to lack of space, focus on strategies with performance assurances that either guarantee success or signal compromised assets.

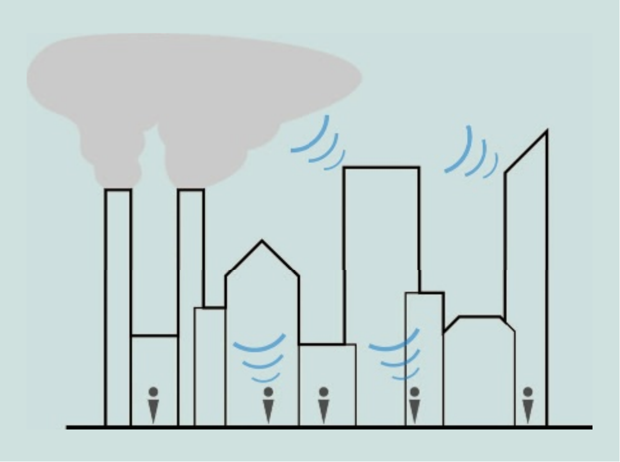


FIGURE 1. One example of an IoT application is air-quality monitoring. In air-quality monitoring, a network of distributed sensors makes measurements of local pollutant concentrations across the city. This information is then processed to produce a heat map of air quality over the city. Users may access this information via smartphones and wearable gadgets; see, e.g., [7].

Preventive and reactive security

To protect data integrity, security countermeasures fall into two main categories: preventive and reactive. Preventive countermeasures seek to prevent intrusion attempts by directly protecting data and communications [11]. Examples of preventive security include cryptographic protocols to authenticate the identity of devices and authorize users to access data. Authentication and authorization protocols ensure that a fusion center receives data streams only from trusted devices. They prevent an adversary from introducing malicious data to the fusion center via a rogue, unverified device [14].

Reactive countermeasures aim to mitigate failures in preventive security and ensure that the system continues to operate properly even when preventive security breaks down [11]. Whereas preventive security protects IoT systems by making it more difficult for an adversary to compromise data and devices, reactive security ensures that systems operate resiliently even when a number of devices become hijacked. Reactive countermeasures include attack detection [15], [16] and identification [17] algorithms for cyberphysical systems (CPSs). The objectives of attack detection and identification are, respectively, to determine if the measurements from any of the sensors have been altered by an adversary and to identify specifically which sensors have been compromised. After detecting or identifying an attack, the system may take corrective action to mitigate the damage [17].

Remark

CPSs refer to physical systems instrumented with a layer of cyberdevices. Examples of CPSs include robotic platforms, drones, and modern automobiles. These are physical systems that are highly instrumented with sensors and actuators (e.g., devices to measure and control speed and acceleration) [9]. At the other extreme of scale in CPSs, large infrastructures, like the power grid, are also highly instrumented by sensors and actuators (e.g., PMUs, smart meters, circuit breakers, and generation sources in the power grid). These devices provide the CPS, be it the modern automobile or the power grid, with a cyberlayer. The devices are not necessarily themselves interconnected, but their measurements are often processed by a central processing unit (CPU), like in an automobile, or by a supervisory control and data acquisition (SCADA) center, like in the power grid. The important characteristic of CPSs is that there is an underlying physical system (e.g., an automobile or the power grid) instrumented by a cyberlayer.

An IoT is more generally a panoply of devices instrumenting, for example, the refrigerator, oven, lighting fixtures, and other appliances in a smart kitchen or connecting wearables, smartphones, and smart speakers to a television [1]. In the IoT, the network provides the ability for the heterogeneous devices to cooperate to achieve a common task. For example, a smart speaker deciphers voice commands and relays them to television to show a specific video, and, of course, in the smart grid, the IoT devices monitor the state of the grid. ■

We further classify reactive countermeasures as either explicit or implicit. Explicit countermeasures directly detect and identify malicious behavior, to alert the system to compensate for adversarial activity. For example, [17] designs an attack identification algorithm for CPSs. In a CPS, such as a remotely controlled vehicle, we are interested in estimating the state of the system (e.g., position, velocity, etc.) from its onboard sensor measurements [e.g., odometer, accelerometer, global positioning system (GPS), camera, lidar]. An adversary may mislead the state estimator by altering some of the sensor measurements [9]. Reference [17] provides a method to identify the compromised sensors and determine the amount by which the measurements were altered. Attack detection and identification, by themselves, may not mitigate the attack. Still, they are important components of secure data processing; once an attack is correctly detected and identified, the IoT system can take corrective actions. For example, in state estimation [17], once the attack has been identified, the state estimator can compensate for the altered measurements and recover the system state.

Implicit countermeasures do not alert IoT systems to malicious behavior. Instead, they provide resilience by limiting the impact of malicious behavior on the system's end goal. As an example, consider a network of sensors monitoring air quality. Individual sensors maintain estimates of global pollutant concentration levels using their measurements of local pollutant concentrations and exchanging estimates with neighboring sensors. Sensors update their estimates as a weighted average of their own previous estimates, their neighbors' previous estimates, and their local measurements. To limit the impact of adversely corrupted measurements, a sensor assigns lower weight to local measurements that deviate more from its estimate [18]. This method may not identify which sensors have been attacked but nevertheless provides resilience against adversarial devices and ensures that the remaining sensors successfully estimate global pollutant concentrations.

IoT architectures and security countermeasures

Countermeasure techniques against data-manipulating adversaries depend on the architecture of the IoT application. In particular, they depend on whether or not the architecture contains a central processor. In centralized architectures, a single processor, possibly in the cloud, has access to data from all devices. For example, in a centralized air-quality monitoring system, the central estimator has access to local measurements of pollutant concentrations from all sensors placed throughout a city and fuses the data to produce a heat map of air quality over the entire city. In this architecture, the individual devices perform minimal processing, and there is no device-to-device communication.

The growing intelligence of IoT devices enables edge, fog, or microedge computing, where some of the computational burden is offloaded from the central processor to the end devices [19]. Like a centralized architecture, a decentralized architecture also features a central processor (i.e., a fusion center) that collects information from individual devices. In decentralized architectures, there is no device-to-device communication. The difference between centralized and decentralized architectures is that, in a decentralized edge architecture, individual devices may process the local data or produce local decisions that are then transmitted to the fusion center [13]. For example, a sensor may transmit a time-moving averaged, quantized version of its data to the fusion center [20]. Decentralized architectures may also be referred to as *parallel architectures*; see Figure 2(a).

In addition to fog or microedge computing, IoT applications will increasingly involve device-to-device communication: that is, instead of transmitting and receiving data to and from the cloud or edge, end devices may communicate directly with each other. The combination of device computing and device-to-device communications enables fully distributed web-like IoT architectures [Figure 2(b)]. We will assume in this article that, in a fully distributed architecture, there is no fusion center and end devices carry all of the computational burden. Devices make local measurements, exchange information with neighboring devices, and process their available information to perform inference. In distributed air-quality monitoring, each sensor will update its estimate pollutant concentrations and communicate this information to neighboring sensors. Through this cooperation, users' end devices obtain air quality about the city, as we will explain when we discuss distributed estimation.

Edge, fog, or microedge computing architectures are better suited than cloud architectures for applications with low latency and real-time processing requirements [19]. Furthermore, the distributed architecture is well suited when there is no central coordinator. One example is vehicular networks in which individual vehicles may exchange information with other nearby vehicles to determine traffic information for path planning or

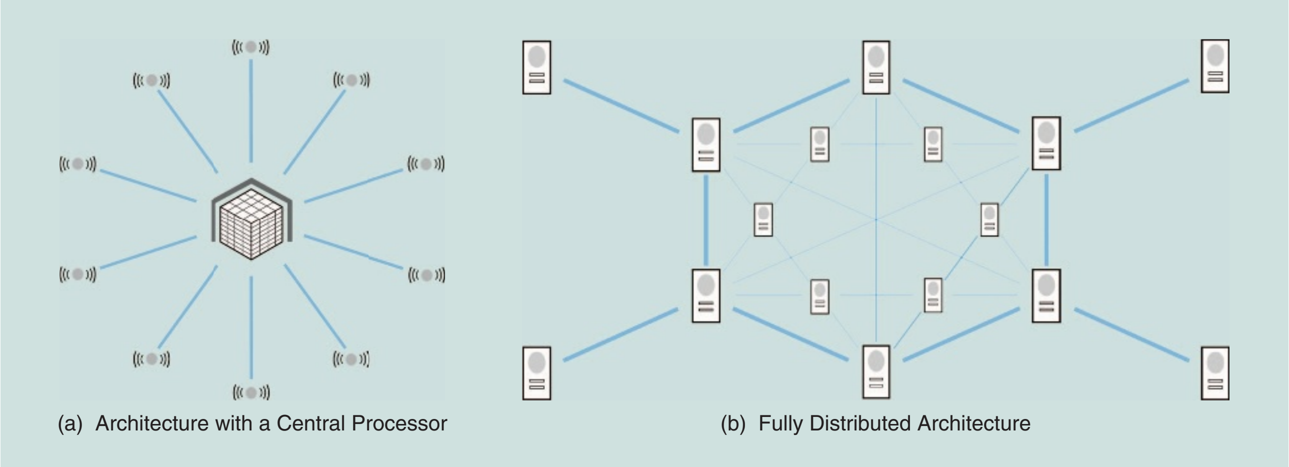

(a) Architecture with a Central Processor
(b) Fully Distributed Architecture

FIGURE 2. (a) Devices transmit raw or processed data to a central processor in centralized and decentralized (or parallel) architectures (see, e.g., [13], [15]–[17], [20], and [21]). There is no device-to-device communication. (b) In distributed architectures, devices communicate with each other and cooperate to complete inference tasks without a fusion center (see, e.g., [12], [18], and [22]–[25]).

to coordinate merging and lane changes [2]. To accommodate the dynamic environment, vehicular driving is computed in real time, so latency looms as a large issue, and computations should be performed at the vehicle rather than in the cloud.

The combination of preventive and reactive countermeasures is suitable for cloud-based IoT architectures with central processors. The authentication and encryption algorithms require high computational power, and the centralized resilient inference algorithms require access to data from all connected devices, both of which are available at a data fusion center. Compared to architectures with central processors, security (in particular, data integrity) in distributed architectures faces more difficult challenges [12]. Individual devices lack the computational capabilities of a cloud data center and may not be able to implement all of the same preventive security measures. In architectures with central processors, preventive countermeasures aim to secure a single entity. In comparison, a distributed architecture consists of numerous devices that are deployed in many different physical locations. Thus, it may be infeasible for preventive security measures to protect all of the devices. Moreover, in a distributed architecture, individual devices do not have access to data from all other devices, and, as a result, they are not able to execute the centralized resilient inference algorithms used in fusion centers.

General measurement and attack model

We model an IoT system as a collection of agents or devices $\{1, 2, \ldots, N\}$, each measuring an unknown parameter θ^*. Individual devices make local, noisy measurements of the phenomenon of interest [15], [20], [25–29]. The measurement $y_n(t)$ of the nth agent is

$$y_n(t) = f_n(\theta^*) + n_t, \qquad (1)$$

where t is time, f_n is the local measurement function, and n_t is the measurement noise. The function f_n is nonlinear in general. An example of nonlinearity is sensor saturation; physical sensors have maximum and minimum measurement levels and exhibit saturation and clipping when the parameter of interest, θ^*, exceeds these bounds. The agents' goal is to recover the parameter θ^* from their measurement streams.

An adversary may hijack a subset of the agents and manipulate their data streams. A standard classical motivation for data integrity attacks against IoT systems comes from the Byzantine generals problem, where a group of generals decides whether or not to attack a city by passing messages among one another (in an all-to-all manner) [30]. Traitor generals attempt to mislead the remaining loyal generals by sending false messages. The authors of [30] provide an algorithm for the loyal generals to reach the correct decision (using an all-to-all communication setup) even when up to one third of the generals are traitors. The classical paradigm of all-to-all communication has been relaxed in, for example, decentralized setups where a fusion center combines all received local decisions (including those of the traitors). In IoT applications, a Byzantine adversary (or Byzantine device) is a device that attempts to disrupt an inference task by transmitting falsified data.

Secure inference with a central processor

State estimation in CPSs

The IoT can monitor critical infrastructure and CPSs, such as the electricity grid and autonomous vehicles. Electricity meters measure power consumption levels and operating conditions of the smart grid [5]. A vital task in operating the smart grid is state estimation, i.e., determining the voltage angles and magnitudes at each bus in the grid from meter measurements. Reference [15] studies state estimation for power grids when an adversary manipulates a subset of the measurements. In [15], a fusion center collects measurements from all of the meters, following the linearized measurement model

$$\mathbf{y} = \mathbf{H}\theta^* + \mathbf{w} + \mathbf{a}. \qquad (2)$$

In (2), $\mathbf{y}$ is the collection of measurements from the meters (each component of $\mathbf{y}$ is a measurement from a single meter), θ^* is the state of the grid (i.e., the voltage angles and magnitudes at each of the buses), $\mathbf{H}$ is a matrix that describes what each meter measures, $\mathbf{w}$ is the measurement noise, and $\mathbf{a}$ models the adversarial attack. In this section, we consider a centralized architecture where a central processor has access to measurements from all of the sensors.

The adversary manipulates the data from a subset of the meters, and the components of $\mathbf{a}$ describe the amount by which a particular measurement is changed. If a meter is not under attack, then the corresponding component of $\mathbf{a}$ is zero. If only a few sensors are attacked, $\mathbf{a}$ is a sparse vector whose nonzero entries may have arbitrary values as determined by the adversary. The goal in state estimation is to recover the value of θ^* from the measurement $\mathbf{y}$. To deal with the adversary, [15] proposes an attack detector. First, the system solves the optimization problem

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \| \mathbf{y} - \mathbf{H}\theta \|_2, \tag{3}$$

that is, the system finds the estimate $\hat{\theta}$ that minimizes the squared error between the measurement $\mathbf{y}$ and the predicted measurement $\mathbf{H}\hat{\theta}$. Then, the attack detector declares that an attack has occurred if the energy of the estimation residual $\mathbf{y} - \mathbf{H}\hat{\theta}$ exceeds a threshold τ, i.e., if $\| \mathbf{y} - \mathbf{H}\hat{\theta} \|_2 > \tau$. This algorithm fails to detect certain attacks. As the authors of [15] show, the adversary can avoid being detected by compromising meters in a specific way such that $\mathbf{a}$ belongs to the column space of the measurement matrix $\mathbf{H}$.

Remark

Resilience in parameter estimation tasks depends on observability. A model is observable if, in the absence of noise and attacks, it is possible to exactly recover the parameter of interest from all the sensor measurements; otherwise, it is unobservable. For the model (2) and centralized architectures, observability means that the matrix $\mathbf{H}$ has full column rank, so there is a unique value of the parameter θ^* that corresponds to any value of the noiseless measurement $\mathbf{H}\theta^*$. We will discuss the notion of observability for distributed architectures in the sequel. The algorithm in [15] detects attacks on up to s sensors as long as the measurement model (2) is observable after removing any s sensors. ■

Reference [15] addresses detecting measurement attacks against static state estimation, i.e., estimating a parameter that does not change over time. For CPSs such as autonomous vehicles, we are interested in estimating a dynamic parameter that changes over time. For example, in context of the measurement model (2), for an autonomous vehicle, the state θ_t describes its position and velocity at time t, and θ_t evolves over time depending on the vehicle's physics. Onboard sensors, such as odometers and GPS, measure the state θ_t and communicate the data to the vehicle's data fusion center. The goal of the fusion center is to recover θ_t, the vehicle's current position and velocity, from the sensor measurements. Just like in the power grid, an adversary can alter the measurement data from some of the vehicle's sensors.

Reference [16] proposes for dynamic CPSs, like autonomous robots, a detector for attacks on sensor measurements. The dynamic attack detector in [16] is similar to the static detector in [15]. The key difference is that [16] uses a model that accounts for a system's dynamics (e.g., laws of physics that describe the motion of a vehicle and bounds on its acceleration) and maps a system state to a sequence of predicted measurements over time. Following this model, the detector from [16] collects a sequence of sensor measurements over time, computes a state estimate, and reports an attack if the energy of the estimate residual (the difference between the observed sensor measurements and the sensor measurements predicted from the state estimate) exceeds a certain threshold. This algorithm detects all sensor attacks so long as the system with only the uncompromised sensors is observable.

The authors of [17] go beyond attack detection and provide an algorithm to identify the sensors under attack. The attack identification algorithm is applicable to both static and dynamic settings. In the context of power grid state estimation (2), the goal of attack identification is to recover the value of the attack $\mathbf{a}$ using the meter measurements $\mathbf{y}$ and the measurement matrix $\mathbf{H}$. The authors of [17] formulate the attack identification problem (in a static setting) as solving the optimization problem

$$\tilde{\theta} = \underset{\theta}{\operatorname{argmin}} \| \mathbf{y} - \mathbf{H}\theta \|_0 \tag{4}$$

and recovering the attack $\mathbf{a}$ as $\tilde{\mathbf{a}} = \mathbf{y} - \mathbf{H}\tilde{\theta}$. Reference [17] assumes a noiseless measurement model, which means, in the context of (2), that $\mathbf{w} = 0$. The idea behind (4) is that the adversary can change the measurements on only a few sensors, so the corresponding attack vector $\mathbf{a}$ contains mostly zeros with sparse nonzero elements. The attack identification algorithm finds the state $\tilde{\theta}$ and the most sparse attack $\tilde{\mathbf{a}}$ that explains the measurement $\mathbf{y}$.

In (4), the goal is to find the estimate $\tilde{\theta}$ that is consistent with the highest number of sensor measurements. The amount by which the observed and predicted measurements (from $\tilde{\theta}$) differ does not matter, since it is assumed that this difference comes as a result of adversarial attack. In contrast, in (3), the goal is to find the estimate that minimizes the total squared error between the observed measurement and the predicted measurement. The algorithm in [17] identifies any attack on up to s sensors if the measurement model (2) is observable after removing any $2s$ sensors. The optimization problem in (4) is nonconvex, and, to solve (4), we must check all possible sets of attacked sensors [17]. The number of possible sets of attacked sensors increases exponentially as the total number of sensors increases. To make the problem tractable, [17] relaxes (4) by replacing the ℓ_0 pseudo-norm with the ℓ_1 norm.

The attack detection [15], [16] and identification [17] algorithms for CPSs are explicit countermeasures against attackers. Their objective is to alert the system to attacks against sensors so that it can take corrective actions to mitigate the effects of the attacks. It is easier to explicitly detect intrusions than it is to design resilient estimation algorithms. The drawback is that attack detection algorithms are incomplete solutions for secure inference since, once an attack has been detected, the system still needs to take corrective action to mitigate the effects of the attack.

Decentralized hypothesis testing

In decentralized hypothesis testing, a group of N sensors measures a phenomenon that falls under one of two hypotheses: $\mathcal{H}_0$ and $\mathcal{H}_1$, occurring with prior probabilities P_0 and P_1, respectively. The sensors communicate with a fusion center whose goal is to determine which hypothesis is true. For example, the sensors can measure environmental conditions in a factory [10], and the goal of the fusion center is to determine whether the conditions are safe $(\mathcal{H}_0)$ or hazardous $(\mathcal{H}_1)$ for the factory's workers. Due to communication constraints, the sensors do not transmit their measurements directly to the fusion center. Instead, each sensor decides, based on its local measurements, which hypothesis is true and transmits the local decision ($\mathcal{H}_0$ or $\mathcal{H}_1$) to the fusion center. In the absence of attacks, the decentralized hypothesis testing problem has been extensively studied [21], [31], [32].

In the presence of Byzantine attacks, an adversary compromises a fraction α of the sensors, and the Byzantine sensors transmit arbitrary decisions to mislead the fusion center. The authors of [26] determine the minimum fraction α of Byzantine agents to ensure that the fusion center cannot distinguish between the two hypotheses. For vector observations, the adversary must compromise at least half of the sensors to ensure that the two hypotheses are indistinguishable [26]. When the fraction of Byzantine sensors is less than one half, [27] designs a fusion rule that is resilient to Byzantines: the fusion center declares hypothesis $\mathcal{H}_1$ if at least K^* of the sensors declare $\mathcal{H}_1$ locally. The threshold K^* depends on the prior probabilities of $\mathcal{H}_0$ and $\mathcal{H}_1$ and the desired level of resilience, i.e., the fraction of Byzantines that needs to be tolerated. Reference [33] analyzes the effect of Byzantine agents in decentralized hypothesis testing in the context of collaborative spectrum sensing.

Security with a central processor: Summary and other work

In architectures with central processors, devices transmit local raw data or local decisions to the central processor, and adversarial devices transmit falsified data to disrupt the inference task. Both explicit and implicit countermeasures require that the uncompromised devices have enough influence to overcome the effects of adversarial behavior. For example, in CPSs, the collection of uncompromised sensors must be observable to detect sensor attacks [16], and, in [27], a majority of devices need to remain uncompromised for the fusion center to resiliently perform hypothesis testing.

Additional work in secure inference with central processors includes [20], which provides an algorithm for resilient decentralized parameter estimation with quantized data. The authors of [28] design methods to identify Byzantine devices in hypothesis testing. Reference [34] surveys attacker strategies and detection methods for data integrity attacks against the smart grid. In addition, [34] proposes a method to detect attacks in the shortest amount of time (i.e., quickest detection). Reference [35] studies state estimation under jamming attacks: in jamming attacks, instead of manipulating data streams, the attacker prevents the devices from communicating with the central processor. The authors of [35] analyze jamming attacks against state estimators in a game-theoretic framework and find Nash equilibrium strategies for both the attacker and the estimator.

Secure inference in distributed architectures

In a fully distributed architecture, there is no fusion center to collect data from all of the devices. This is a simplistic architecture, since in the real world we may expect a hybrid or hierarchical architecture where devices communicate among themselves as well as with the edge (intermediate computing resources) and the cloud. At different levels, different strategies may be used, including a mix of the ones we described in the section "Secure Inference with a Central Processor" and the ones we consider here. In a distributed architecture, devices communicate with each other to complete computation and inference tasks.

For simplicity, we consider a flat network of N devices (or agents), $\{1, 2, \ldots, N\}$. We model the communication between devices with an undirected simple graph $G = (V, E)$. For background on graphs, see [36]. The vertex set V of G is the set of N devices, and the edge set E describes the communication links among them. Two devices are connected by an edge if they can communicate with each other. A device can only communicate with its neighbors in the graph G. The set Ω_n is the neighborhood of device n, i.e., the set of all devices that share a communication channel with device n. For example, for distributed air-quality monitoring in a city, individual sensors may only communicate with nearby sensors, instead of communicating with all other sensors in the city. We now consider two distributed inference tasks. The first is distributed consensus, where sensors or devices cooperatively compute a statistic of a snapshot of data, for example, the average of the (distributed) data. The second is distributed estimation, where sensors cooperate iteratively to process a stream of measurements and recover the value of an unknown parameter.

Resilient consensus

In consensus, a network of devices cooperates to agree on a common value [22]–[24]. Consensus is important in distributed IoT architectures because it ensures that, in computation or inference tasks, all devices agree on the result. Reference [23] studies distributed average consensus: each of the N devices is assigned an initial scalar value, and their goal is to compute the average value of all of the devices. In an air-quality monitoring application, a network of sensors could cooperate to find the average pollutant concentration in a city. Every device n maintains a local scalar value $x_n(t)$, where t is an iteration, with

$x_n(0)$ equal to its initial assigned value. Again, in the air-quality monitoring example, $x_n(0)$ represents the local pollutant concentration at sensor n. Then, every device n transmits its current value or state $x_n(t)$ to all of its neighboring devices (device n also receives from all its neighbors their current states) and updates its state as a weighted sum of its current state and its neighbors states

$$x_n(t+1) = w_{nn}x_n(t) + \sum_{j \in \Omega_n} w_{jn}x_j(t). \tag{5}$$

For properly chosen weights $\{w_{jn}\}_{j,n}$, the states at each device converge toward the average of the initial data [23].

An adversary may hijack certain devices and disrupt the consensus process. Reference [37] studies distributed consensus when some devices follow an update rule that deviates from (5). That is, compromised devices do not follow (5) when updating their values and instead update their values arbitrarily. To counter this adversarial behavior, the authors of [37] design algorithms to detect and identify compromised devices. Recall that attack detection and identification algorithms are explicit countermeasures against adversaries. The main idea in [37] is to model the consensus process as a linear dynamical system and view attacks from compromised agents as unknown inputs into the system. To detect compromised devices, an agent must determine if there is a nonzero unknown input into the system; to identify compromised devices, an agent must determine the locations of the nonzero attacks. From these ideas, [37] designs algorithms for each agent n to detect and identify other compromised agents using only the history of its own $x_n(t)$. These algorithms require each device to store the topology of the communication network, which becomes computationally infeasible as the number of devices in the system grows, as with IoT applications.

Reference [38] designs an implicit countermeasure against adversaries in consensus. In [38], instead of computing the average of their initial values, the devices' goal is to simply agree on a value. That is, the devices wish to update their states such that, eventually, the uncompromised devices reach the same value. The compromised devices update their states arbitrarily and transmit falsified values to their neighbors to disrupt consensus. The authors of [38] modify the state update rule (5) to deal with the compromised devices. When an agent updates its state, instead of using all of its neighbors' states, it ignores the most extreme state values. Before updating, each device n sorts the states received from its neighbors Ω_n and removes the largest F state values greater than its own and the smallest F state values lower than its own, for some predetermined number F. Then, each agent n updates its state $x_n(t+1)$ as a weighted sum of the states from its remaining neighbors and its own current state $x_n(t)$. As long as the total number of compromised devices is fewer than F and the communication network satisfies certain topology conditions, the modified state update rule ensures that all uncompromised agents consensus on the same state.

Secure parameter estimation: Implicit countermeasures

The section "Resilient Consensus" focused on distributed consensus, where devices converge to a common statistic from a single snapshot of their data, e.g., the average of their initial data. In distributed inference, for example, like distributed estimation, devices still converge to a common estimate of an unknown parameter, but at each communication round they make a new observation. We present three implicit countermeasures ([18], [29], and [39]) for secure distributed inference. Reference [39] extends the resilient consensus algorithm in [38] to construct a resilient distributed estimator. In [39], each device n processes a stream of local measurements to recover a scalar local parameter p_n. In the context of air-quality monitoring, the parameter p_n could be the concentration of pollutants at sensor n.

The authors in [39] consider three different types of devices to be part of the network: 1) reliable, 2) normal, and 3) malicious. Reliable devices directly measure their parameters of interest, and they can recover their parameters using only their local data. Normal devices are not able to directly measure their parameter. Instead, a normal device n makes a relative measurement $\xi_{ln}(t) = p_l - p_n$ for every neighboring device l. Each device n maintains an estimate $x_n(t)$, i.e., its state, of its local parameter p_n. At every time step, each device broadcasts its state to all its neighbors. Malicious devices may broadcast an arbitrary state or estimate. Reliable devices measure their parameter directly, so their state is the correct estimate $x_n(t) = p_n$ for all iterations. Each normal device n, for each of its neighbors $l \in \Omega_n$, computes a step state value $s_{ln}(t) = x_l(t) - x_n(t) - \xi_{ln}(t)$. Device n ignores the largest F positive step values and the smallest F negative step values. If there are fewer than F positive (negative) step values, then device n ignores all positive (negative) step values. Then, each device n updates its state as a weighted sum of its previous state and the remaining estimates. Reference [39] shows that if there are fewer than F malicious devices in any device's neighborhood and if the topology of the communication network satisfies certain conditions, then, all reliable and normal devices eventually recover their parameters.

Another method for devices to deal with adversaries is to apply different gains or weights to their measurements and the information they receive from neighbors [18], [29]. In [29], a network of devices make noisy measurements of an underlying parameter. The devices maintain local estimates of the parameter and update them as a weighted sum of their previous estimates (states), the states of their neighbors, and their local measurement. Malicious devices attempt to disrupt the estimation process by broadcasting false estimates to their neighbors. The authors of [29] propose an adaptive weight estimate update scheme, where an uncompromised device gives lower weight to neighbors whose estimates deviate drastically from its own. Through this adaptive weight mechanism, the uncompromised devices learn to eventually ignore malicious devices and, effectively, disconnect the adversaries from the network.

In [18], instead of hijacked devices broadcasting false estimates, the adversary attacks the network by manipulating the devices' measurements. In air-quality monitoring, this corresponds to the case in which an attacker falsifies the sensor data (say, the measurement of local pollutant concentrations) on a subset of devices. The devices all observe the complete parameter and apply an adaptive gain to their measurements to mitigate the effect of the attack—each device n gives lower weight to local measurements that deviate more from its local estimate. Reference [18] shows that applying lower weights to aberrant measurements limits the impact maliciously altered measurements. If fewer than half of the devices fall under attack, the network eventually recovers correctly the parameter of interest.

Secure distributed inference: Other work

Further work on secure distributed inference includes inference under jamming attacks [40] and function calculation [41], distributed hypothesis testing [42], [43], and distributed optimization [44]–[47] with Byzantine agents. Reference [40] studies distributed estimation under jamming attacks: in jamming attacks, the adversary prevents communication between devices instead of manipulating their data streams. In [41], the authors design an algorithm that is resilient to Byzantines for computing a specific function of a single snapshot of data (this differs from consensus, where the goal is for the agents to converge to any common statistic). Reference [42] studies distributed hypothesis testing with Byzantines and provides an algorithm that is resilient to a restricted class of weak Byzantine adversaries. The authors of [43] evaluate a heuristic for Byzantine-resilient distributed hypothesis testing through numerical simulations.

In distributed optimization, each agent has a local objective function, and the agents' goal is to converge to a statistic that minimizes the sum of their objective functions, possibly subject to constraints. Reference [44] considers optimization in an all-to-all communication setup (i.e., each agent communicates with every other agent) and proposes an iterative optimization algorithm that is resilient to Byzantine agents. References [45] and [46] present optimization algorithms that are resilient to Byzantine agents for arbitrary network topologies. The authors of [47] present a Byzantine-resilient distributed optimization algorithm for training a support vector machine.

Secure distributed estimation through explicit adversary detection

The algorithms provided in [18], [29], and [39] are implicit security countermeasures for distributed estimation: they ensure that a network of devices completes the estimation task even when adversaries attack the network, but they do not detect or identify the adversaries. We now consider an explicit countermeasure. Reference [25] designs an algorithm for the subset of uncompromised devices in the network, the ones not under attack, to still simultaneously infer the value of a parameter from their stream of measurements or detect the presence of compromised devices. In a sense, this is a 0–1 strategy, the sensors surviving the attack still achieve the desired goal, or, if the attack is too strong, they are able to detect the attack and realize the presence of an intruder.

Device model

Each device n makes a stream of measurements, $y_n(t)$, of a parameter θ^* following

$$\mathbf{y}_n(t) = \mathbf{H}_n \theta^* + \mathbf{w}_n(t), \tag{6}$$

where $w_n(t)$ is measurement noise and the matrix $\mathbf{H}_n$ describes which parts of the parameter each device measures. For example, in air-quality monitoring, θ^* represents the pollutant concentration over an entire city, and each individual component of θ^* may represent the pollutant concentration in a particular neighborhood. At each device n, the matrix $\mathbf{H}_n$ selects the component of θ^* corresponding to the local pollutant concentration. The parameter θ^* has bounded energy (i.e., $\|\theta^*\|_2 \leq \eta$ for some known constant η), since, in practice, we are interested in parameters bound by physical laws. Again, in air-quality monitoring, by definition, no pollutant concentration can be above 10^6 parts per million, and, in practice, the bound may be even tighter. The noise $\mathbf{w}_n(t)$ is independently and identically distributed (i.i.d.) with mean $\mathbb{E}[\mathbf{w}_n(t)] = 0$ and finite covariance $\mathbb{E}[\mathbf{w}_n(t)\,\mathbf{w}_n(t)^\top] = \Sigma_n$ and is independent across devices.

The network of devices, $G = (V,E)$, satisfies two natural conditions. First, the graph G is connected. There is a path between any two devices, and information from each device propagates to all other devices. Second, the network of devices is globally observable: the matrix $\Sigma_{n=1}^{N} \mathbf{H}_n^\top \mathbf{H}_n$ is invertible, where the matrices $\mathbf{H}_n$ model the local measurement at sensor n in (6). This global observability condition is equivalent to the rank observability conditions for centralized architectures. Intuitively, global observability means that the sensors together provide meaningful information about each component of θ^*. In air-quality monitoring, global observability means loosely that the collective of all devices provides information about pollutant concentrations in all neighborhoods, but each individual sensor needs to measure only the local pollutant concentration of a neighborhood.

The goal is to recover the parameter θ^* from the measurements of the networked devices. This is to be achieved through cooperation among the devices, with each device iteratively updating its local estimate and broadcasting this to its neighbors. We assume that a subset, $\mathcal{A}$, of the devices are Byzantine, and they broadcast arbitrary estimates to their neighbors. The Byzantine devices are the same in each time step, i.e., the set $\mathcal{A}$ does not change over time. The remaining uncompromised devices, $\mathcal{N}$, wish to recover θ^* even in the presence of malicious attacks of the devices in $\mathcal{A}$.

Distributed estimation with local consistency checks

We now describe a resilient distributed estimation algorithm. The resilient algorithm combines a distributed detection step with distributed estimation. If the detector does not detect the presence of Byzantine actors, then, the estimation step converges with probability one to the correct value θ^*, even if a subset of the agents is compromised. At each time step $t = 0, 1, 2, \ldots$, agent n maintains a local estimate or state $\mathbf{x}_n(t)$ and a flag $\pi_n(t)$. The flag $\pi_n(t)$ is either "Attack" or "No Attack," indicating the presence (or absence) of adversaries. The algorithm iterates among three main steps: 1) message passing, 2) state update, and 3) adversary detection. Each (uncompromised) agent n initializes its state and flag as $x_n(0) = 0$ and $\pi_n(0) =$ No Attack. Compromised agents will act arbitrarily. So, here, we only specify the rules followed by the uncompromised agents. While compromised agents follow the policy described by the attacker, the uncompromised agents adhere to the following rules.

Message passing

At time $t = 0, 1, 2, \dots,$ uncompromised agent $n \in \mathcal{N}$ transmits its current state, $\mathbf{x}_n(t)$, to its neighbors.

State update

To average out the disturbance from the measurement noise $\mathbf{w}_n(t)$, uncompromised agent n maintains a time-running average of its local measurement:

$$\begin{aligned} \bar{\mathbf{y}}_n(t) &= \frac{t}{t+1}\bar{\mathbf{y}}_n(t-1) + \frac{1}{t+1}\mathbf{y}_n(t), \\ \bar{\mathbf{y}}_n(0) &= \mathbf{y}_n(0). \end{aligned} \tag{7}$$

The uncompromised agents $n \in \mathcal{N}$ update their states following a consensus plus innovations rule (see, e.g., [48]):

$$\begin{aligned} \mathbf{x}_n(t+1) = \mathbf{x}_n(t) & \underbrace{- \beta \sum_{l \in \Omega_n} (\mathbf{x}_n(t) - \mathbf{x}_l(t))}_{\text{Consensus}} \\ & + \underbrace{\alpha \mathbf{H}_n^T (\bar{\mathbf{y}}_n(t) - \mathbf{H}_n \mathbf{x}_n(t))}_{\text{Innovations}}. \end{aligned} \tag{8}$$

The weights α and β are positive weighting for the innovations and consensus terms, respectively, in (8), where the innovations term incorporates local measurements and the consensus term propagates local measurements throughout the network and drives the agents to reach the same estimate.

Adversary detection

Each uncompromised agent $n \in \mathcal{N}$ checks for adversaries by comparing its own estimate with the estimates it receives from its neighbors. An agent reports the presence of adversaries if the (Euclidean) distance between its own state and the state from any of its neighbors exceeds an adaptive threshold. The uncompromised agents update their flags following

$$\pi_n(t+1) = \begin{cases} \text{Attack,} & \pi_n(t) = \text{Attack, or} \\ & \exists l, \| x_n(t) - x_l(t) \|_2 > \gamma_t, \\ \text{No Attack,} & \text{Otherwise} \end{cases} \tag{9}$$

where γ_t is a time-varying adaptive threshold. The threshold γ_t follows the recursion

$$\begin{aligned} \gamma_{t+1} &= \underbrace{(1 - r_1)\gamma_t}_{\text{Error Buffer}} + \underbrace{\alpha \frac{2K}{(t+1)^\tau}}_{\text{Noise Buffer}}, \\ \gamma_0 &= 2\eta\sqrt{N}, \end{aligned} \tag{10}$$

and depends on the parameters $K > 0$, $0 < \tau < 1/2$, and $0 < r_1 \le 1$. Recall that η bounds the energy of the parameter θ^*, and N is the total number of devices.

The threshold γ_t describes how far apart the states of two neighboring devices should be if there is no adversary. It consists of two components: the error buffer component, $(1 - r_1)\gamma_t$, and the noise buffer component, $\alpha(2K)/((t+1)^\tau)$. As the agents follow the state update, their states move closer to those of their neighbors. The error buffer describes the rate at which neighboring devices' states move closer together in the absence of adversaries and noise. The noise buffer compensates for the effect of measurement noise and depends on the parameters K and τ. The parameter K describes the base size of the noise buffer at each iteration, and the parameter τ describes how the noise buffer decays over time. If the threshold γ_t is too small (e.g., if base size K of the noise buffer is too small), then the algorithm incurs a high probability of false alarm, since measurement noise may cause neighboring agents' states to exceed threshold.

Adversarial devices update their own estimates in an arbitrary manner and have no need for a flag. To avoid detection, adversarial agents, which, in the extreme case we assume, know all algorithm parameters, must ensure that, for all times t, the distance between the state they transmit and each of their neighbors' states is lower than the threshold γ_t. There is a tradeoff between the magnitude of the threshold and the performance of the algorithm. For large threshold values, the algorithm has few false alarms, but adversarial devices may transmit estimates that deviate more significantly while evading detection. Small thresholds detect adversaries more effectively but suffer from more false alarms.

The beauty of the approach in [25] is that, by careful design of the algorithm parameters α, β, and γ_t, one can guarantee (see the section "Estimator Performance") that either an attack is detected or the estimator is accurate. The parameters K, the base size of the noise buffer, and τ, the decay rate of the noise buffer, may take any values that satisfy $K > 0$ and $0 < \tau < 1/2$. For the agents to effectively recover θ^* and detect adversaries with low false alarm probability, we must choose α, the innovation weight, β, the consensus weight, and r_1, the decay rate of the error buffer, to satisfy certain eigenvalue conditions related to the dynamics of the estimate update rule (8). The algorithm in [25] is fully distributed and requires only local data at each agent. This differs from attack detectors for architectures with central processors (e.g., [15]), which require the central processor to have access to all data streams. Additionally, the algorithm in [25] does not require each agent to store the topology of the network locally, unlike the attack detector for consensus algorithms presented in [37].

Estimator performance

In the absence of Byzantine agents, the algorithm from [25] ensures that all agents produce strongly consistent estimates (i.e., they eventually recover the parameter θ^* almost surely) and have few false alarms. The false alarm rate can be made arbitrarily small by choosing a larger noise buffer base size K. When there are Byzantine agents, the performance of the algorithm depends on the distributed observability of the remaining, uncompromised agents, $\mathcal{N}$. Consider the network of uncompromised agents only, and suppose that this network is connected and globally observable. In the presence of adversarial agents, one of two events must occur: either 1) at some time t an uncompromised agent n changes its flag value to $\pi_n(t) = $ Attack, or 2) no uncompromised agent ever changes its flag value. If the first event occurs, the algorithm successfully detects the Byzantine agents.

If the second event occurs, the adversarial agents evade detection. To evade detection, each adversarial agent n may only transmit states that deviate from their neighbors' states by less than the threshold, γ_t. The adaptive threshold decays over time, which means that, to evade detection, the adversarial agents' attack must become weaker over time. Under the conditions of connectivity and global observability, the network of uncompromised agents still produces consistent estimates when the adversarial agents evade detection. In this sense, the algorithm from [25] outperforms standard anomaly detectors for distributed architectures (e.g., [37]): the distributed attack detector guarantees that, if there is a missed detection, then the agents still produce consistent estimates. Standard anomaly detectors provide no such guarantee. The key conditions for resilience under the algorithm from [25] are that the network of uncompromised devices is connected and globally observable. If these conditions are not satisfied, then it is possible for the adversaries to disrupt the estimation process (i.e., cause the devices to produce inconsistent estimates) while evading detection.

Numerical example

As an illustration, consider air-quality monitoring in smart cities. For example, the city of Chicago plans to deploy 500 sensors by the end of 2018 to monitor environmental conditions [6]. Figure 3 shows a network of $N = 500$ sensors deployed in nine sectors of a city. Sensors are placed uniformly at random over a 2 km × 2 km grid. Two sensors share a communication link if they are located within 200 m of each other. Each sensor measures the pollutant concentration in its own sector only, and their goal is to recover the pollutant concentrations over all nine sectors.

Each component of θ^*, representing local pollutant concentrations, is drawn independently and uniformly between 0 and 160 $\mu g/m^3$. According to the U.S. Environmental Protection Agency, the maximum safe level of particulate matter 10 (PM_{10}) is 150 $\mu g/m^3$ [49]. Each device's sensor is corrupted by additive Gaussian white noise with mean zero and variance $\Sigma_n = 10$. The local signal-to-noise ratio (SNR) is 13 dB. We demonstrate the performance of the algorithm in three different scenarios:

1) *No adversaries*: All devices remain uncompromised.
2) *Strong adversaries*: An adversary compromises all devices in the center sector. The remaining devices are no longer globally observable.
3) *Weak adversaries*: An adversary compromises half of the devices in the center sector. The remaining devices are connected and globally observable.

Figure 4 depicts the performance of the algorithm and shows the evolution of the agents' estimation errors and flag values

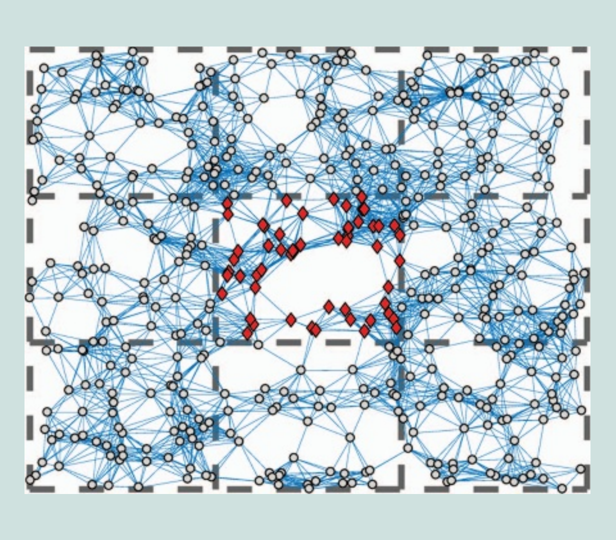


FIGURE 3. Network of $N = 500$ sensors. Each sensor measures pollutant concentrations in its local sector only. An adversary hijacks a subset of sensors in the center sector, denoted by red diamonds.

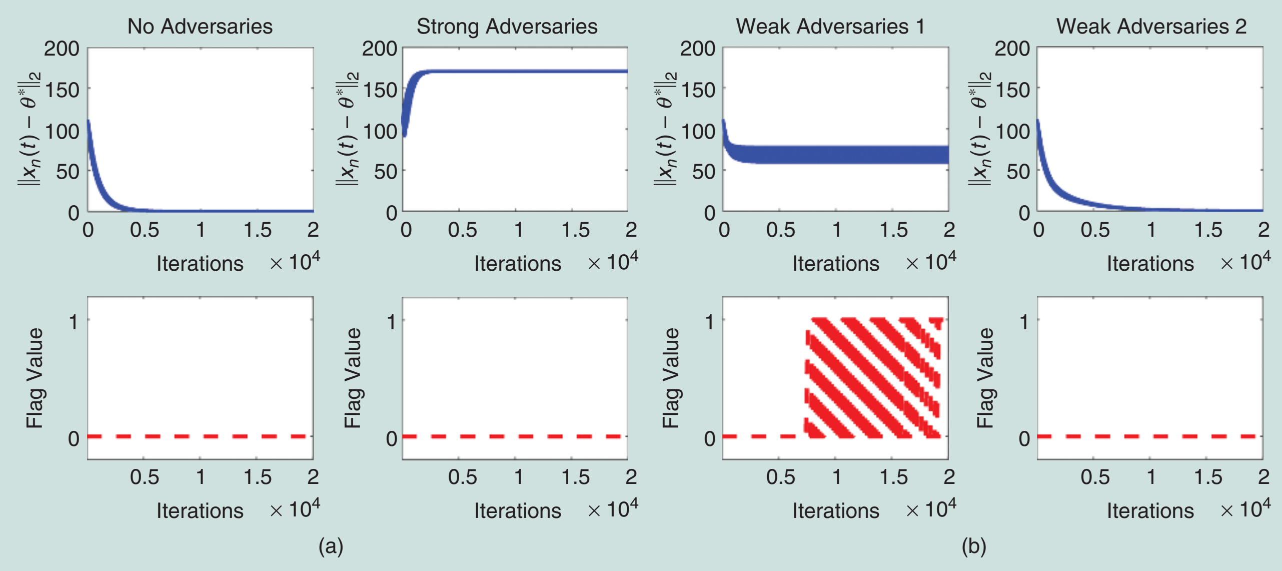


FIGURE 4. The evolution of estimation errors and flag values over iterations of (8) and (9) for uncompromised devices with (a) no adversaries/strong adversaries and (b) weak adversaries. When an agent detects an adversary, it changes its flag value from zero to one.

over 20,000 iterations. When there are no compromised devices, the estimates of all devices converge to θ^*, and no device reports the presence of adversaries. When all of the devices in the center sector are compromised, the remaining devices are unable to recover θ^* and do not detect the adversaries. This is because the network of remaining uncompromised devices is not globally observable. It has no information about the pollutant concentrations in the center sector. When only half of the devices in the sector are compromised, the network of uncompromised devices is globally observable. In this case, adversaries that disrupt the estimation process are eventually detected. If adversaries attempt to evade detection, then the remaining devices eventually recover the global pollutant concentration, θ^*, although, in this case, the devices' estimates converge more slowly compared to the case where there are no adversaries.

Conclusions

In this article, we presented an overview of methods for resilient decentralized and distributed inference in the IoT. We have separately considered explicit countermeasures, such as adversary detection and identification algorithms, and implicit countermeasures, inference algorithms that are inherently resilient to data manipulation. A general requirement for achieving resilience is that the uncompromised, cooperative devices have enough influence to overcome the disruptive effects of adversarial devices. In simple settings, e.g., where all devices observe the same phenomena, this means that a majority of devices should be uncompromised. In settings where devices observe different phenomena, for example, different components of an unknown parameter, resilience depends on the observability of the uncompromised devices.

There are several open challenges for secure distributed inference in the IoT. First, for fully distributed IoT systems, we have focused primarily on static inference tasks, e.g., estimating a parameter that does not change (or changes slowly) over time. It is also necessary to design countermeasures for dynamic distributed inference tasks, where the target parameter changes quickly in time or where agents move and the network changes over time, e.g., a network of automobiles estimating traffic conditions. In cases where agents are mobile, an adversarial agent may move into different agents' neighborhoods over time, making the problem of detecting and identifying adversaries more challenging.

Second, we have focused on inference tasks where all of the devices aim to recover the same parameter or decision. Another area of future work is designing resilient algorithms for inference tasks where devices have different goals. For example, in air-quality monitoring, a device may be interested in recovering the pollutant concentration in its sector and nearby sectors only instead of recovering the pollutant concentrations over an entire city.

Finally, we have focused on countermeasures that ensure systems complete their inference tasks. Depending on the adversary, this may not be possible, e.g., if, in decentralized hypothesis testing, the majority of devices is compromised. A goal of future work is to design countermeasures that ensure graceful performance degradation when complete resilience is not achievable.

Acknowledgments

Figures and diagrams are courtesy of Jasper Tom. This material is based on work supported by the U.S. Department of Energy under Award Number DE-OE0000779 and by the U.S. Defense Advanced Research Projects Agency (DARPA) under agreement numbers DARPA FA8750-12-2-0291 and DARPA HR00111320007. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

Authors

Yuan Chen (yuanchen.yc13@gmail.com) received his B.S.E. degree in electrical engineering from Princeton University, New Jersey, in 2013. Since 2013, he has been pursuing his Ph.D. degree in electrical and computer engineering at Carnegie Mellon University, Pittsburgh, Pennsylvania. His current research activities are focused on distributed inference, cyberphysical systems, and security for the Internet of Things. He is a Student Member of the IEEE.

Soummya Kar (soummyak@andrew.cmu.edu) received his B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, in 2005 and his Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, Pennsylvania, in 2010. From June 2010 to May 2011, he was with the Electrical Engineering Department, Princeton University, New Jersey, as a postdoctoral research associate. He is currently an associate professor of electrical and computer engineering at Carnegie Mellon University. His research interests include decision making in large-scale networked systems, stochastic systems, multiagent systems, and data science, with applications to cyberphysical systems and smart energy systems. He is a Member of the IEEE.

José M.F. Moura (moura@ece.cmu.edu) received the EE degree from the Instituto Superior Técnico, Lisbon, Portugal, and the M.Sc., E.E., and D.Sc. degrees from the Massachusetts Institute of Technology, Cambridge. He is the Philip L. and Marsha Dowd University Professor at Carnegie Mellon University (CMU). His research interests are in data science and graph signal processing. Technology of two of these, invented with A. Kavcic, are found in over 3 billion disk drives in 60% of all computers sold in the last 13 years and led to the US$750 million settlement between CMU and Marvell, the largest ever in the IT area. He has published more than 550 papers and holds 15 patents. He received the Technical Achievement Award and the Society Award from the IEEE Signal Processing Society. He is a Fellow of the IEEE, the American Association for the Advancement of Science, and the U.S. National Academy of Inventors. He is a member of the U.S. National Academy of Engineering and a corresponding member of the Academy of Sciences of Portugal. He is the 2018 IEEE president-elect.

References

[1] M. Rouse, S. Shea, and M. Haughn. IoT devices (Internet of Things devices). [Online]. Available: http://internetofthingsagenda.techtarget.com/definition/IoT-device, Mar., 2018.

[2] N. Lu, N. Cheng, N. Zhang, X. Shen, and J. W. Mark, "Connected vehicles: Solutions and challenges," *IEEE Internet Things J.*, vol. 1, no. 4, pp. 289–299, Aug. 2014.

[3] A. S. El-Wakeel, J. Li, A. Noureldin, H. S. Hassanein, and N. Zorba, "Towards a practical crowdsensing system for road surface conditions monitoring," *IEEE Internet Things J.*, vol. PP, no. 99, pp. 1–14, Feb. 2018.

[4] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: Current state and future challenges," *IEEE Commun. Mag.*, vol. 49, no. 11, pp. 32–39, Nov. 2011.

[5] N. Bui, A. P. Castellani, P. Casari, and M. Zorzi, "The Internet of energy: A web-enabled smart-grid system," *IEEE Netw.*, vol. 26, no. 4, pp. 39–45, July 2012.

[6] What is the array of things? [Online]. Available: http://www.arrayofthings.github.io/faq.html

[7] J. Venkatesh, B. Aksanli, C. S. Chan, A. S. Akyurek, T, and S. Rosing, "Modular and personalized smart health application design in a smart city environment," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 614–623, Apr. 2018.

[8] S. Tan, D. De, W. Song, J. Yang, and S. K. Das, "Survey of security advances in smart grid: A data driven approach," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, Jan. 2017, pp. 397–422.

[9] F. Franchetti, T. M. Low, S. Mitsch, J. P. Mendoza, L. Gui, A. Phaosawasdi, D. Padua, S. Kar, J. M. F. Moura, M. Franusich, J. Johnson, A. Platzer, and M. M. Veloso, "High-assurance SPIRAL: End-to-end guarantees for robot and car control," *IEEE Control Syst. Mag.*, vol. 37, no. 2, pp. 82–103, Apr. 2017.

[10] M. Stolpe, "The Internet of Things: Opportunities and challenges for distributed data analysis," *ACM SIGKDD Explorations Newslett.*, vol. 18, no. 1, pp. 15–34, June 2016.

[11] I. Tomić and J. A. McCann, "A survey of potential security issues in existing wireless sensor network protocols," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 1910—1923, Dec. 2017.

[12] P. Roman, J. Zhou, and J. Lopez, "On the features and challenges of security and privacy in distributed Internet of Things," *Comput. Netw.*, vol. 57, no. 10, pp. 2226–2279, July 2013.

[13] JN. Tsitsiklis, "Decentralized detection," in *Advances in Statistical Signal Processing, Vol. 2: Signal Detection*, H. V. Poor and J. B. Thomas, Eds. Greenwich, CT: JAI Press, 1993, pp. 297–344.

[14] H. Kim and E. A. Lee, "Authentication and authorization for the Internet of Things," *IT Prof.*, vol. 19, no. 5, pp. 27–33, Oct. 2017.

[15] Y. Liu, M. K. Reiter, and P. Ning, "False data injection attacks against power systems in electric power grids," in *Proc. 16th ACM Conf. Computer and Communications Security*, Chicago, IL, Nov. 2009, pp. 21–32.

[16] Y. Chen, S. Kar, and J. M. F. Moura, "Dynamic attack detection in cyber-physical systems with side initial state information," *IEEE Trans. Autom. Control*, vol. 62, no. 9, pp. 1–7, Sept. 2017.

[17] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Trans. Autom. Control*, vol. 59, no. 6, pp. 1454–1467, June 2014.

[18] Y. Chen, S. Kar, J. M, and F. Moura, "Resilient distributed estimation: Sensor attacks," *ArXiv e-prints*, pp. 1–8, Mar. 2018.

[19] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[20] J. Zhang, R. Blum, X. Lu, and D. Conus, "Asymptotically optimum distributed estimation in the presence of attacks," *IEEE Trans. Signal Process.*, vol. 63, no. 5, pp. 1086–1101, Mar. 2015.

[21] J. N. Tsitsiklis, "Decentralized detection by a large number of sensors," *Math. Contr., Signals, System*, vol. 1, no. 2, pp. 167–182, June 1988.

[22] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. 31, no. 9, pp. 803–812, 1986.

[23] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Syst. Control Lett.*, vol. 53, no. 1, pp. 65–78, Sept. 2004.

[24] S. Zhu and B. Chen, "Quantized consensus by ADMM: Probabilistic versus deterministic quantizers," *IEEE Trans. Signal Process.*, vol. 64, no. 7, pp. 1700–1713, Apr. 2016.

[25] Y. Chen, S. Kar, and J. M. F. Moura, "Resilient distributed estimation through adversary detection," *IEEE Trans. Signal Process.*, vol. PP, no. 99, pp. 1–15, Mar. 2018.

[26] S. Marano, V. Matta, and L. Tong, "Distributed detection in the presence of Byzantine attacks," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 16–29, Jan. 2009.

[27] B. Kailkhura, Y. S. Han, S. Brahma, and P. K. Varshney, "Distributed Bayesian detection in the presence of Byzantine data," *IEEE Trans. Signal Process.*, vol. 63, no. 19, pp. 5250–5263, Oct. 2015.

[28] A. Vempaty, L. Tong, and P. K. Varshney, "Distributed inference with Byzantine data," *IEEE Signal Process. Mag.*, vol. 30, no. 5, pp. 65–75, Sept. 2013.

[29] A. H. Sayed, S. Tu, J. Chen, X. Zhao, and Z. J. Towfi, "Diffusion strategies for adaptation and learning over networks," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 155–171, May 2013.

[30] L. Lamport, R. Shostak, and M. Pease, "The Byzantine generals problem," *ACM Trans. Program. Lang. Syst.*, vol. 4, no. 3, pp. 382–401, July 1982.

[31] S. Alhakeem and P. K. Varshney, "A unified approach to design of decentralized detection systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 31, no. 1, pp. 9–20, Jan. 1995.

[32] S. A. Aldosari and J. M. F. Moura, "Detection in sensor networks: The Saddlepoint approximation," *IEEE Trans. Signal Process.*, vol. 55, no. 1, pp. 327–340, Dec. 2006.

[33] A. S. Rawat, P. Anand, H. Chen, and P. K. Varshney, "Collaborative spectrum sensing in the presence of Byzantine attacks in cognitive radio networks," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 774–786, Nov. 2010.

[34] S. Cui, Z. Han, S. Kar, T. T. Kim, H. V. Poor, and A. Tajer, "Coordinated data-injection attack and detection in the smart grid," *IEEE Signal Process. Mag.*, vol. 29, no. 5, pp. 106–115, Sept. 2012.

[35] Y. Li, L. Shi, P. Cheng, J. Chen, and D. E. Quevedo, "Jamming attacks on remote state estimation in cyber-physical systems: A game theoretic approach," *IEEE Trans. Autom. Control*, vol. 60, no. 10, pp. 2831–2836, Oct. 2015.

[36] F. R and K. Chung, *Spectral Graph Theory*. Providence, RI: Wiley, 1997.

[37] F. Pasqualetti, A. Bicchi, and F. Bullo, "Consensus computation in unreliable networks: A system theoretic approach," *IEEE Trans. Autom. Control*, vol. 57, no. 1, pp. 90–104, Jan. 2012.

[38] H. J. LeBlanc, H. Zhang, X. Koustsoukos, and S. Sundaram, "Resilient asymptotic consensus in robust networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 4, pp. 766–781, Apr. 2015.

[39] H. J. LeBlanc and F. Hassan, "Resilient distributed parameter estimation in heterogeneous time-varying networks," in *Proc. 3rd Int. Conf. High Confidence Networked Systems*, Berlin, Germany, Apr. 2014, pp. 19–28.

[40] Y. Guan and X. Ge, "Distributed secure estimation over wireless sensor networks against random multichannel jamming attacks," *IEEE Access*, vol. 5, pp. 10858–10870, June 2017.

[41] S. Sundaram and C. N. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Trans. Autom. Control*, vol. 56, no. 7, pp. 1495–1508, July 2011.

[42] B. Kailkhura, S. Brahma, and P. K. Varshney, "Data falsification attacks on consensus-based detection systems," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 1, pp. 145–158, Sept. 2016.

[43] B. Kailkhura, P. Ray, D. Rajan, A. Yen, P. Barnes, and R. Goldhahn, "Byzantine-resilient locally optimum detection using collaborative autonomous networks," in *Proc. IEEE Int. Workshop Computational Advances Multi-Sensor Adaptive Processing*, Curaçao, Dutch Antilles, Dec. 2017, pp. 1–5.

[44] L. Su and N. H. Vaidya, "Fault-tolerant multi-agent optimization: Optimal iterative distributed algorithms," in *Proc. ACM Symp. Principles Distributed Computing*, Chicago, IL, July 2016, pp. 425–434.

[45] Z. Yang, W and U. Bajwa, "ByRDiE: Byzantine-resilient distributed coordinate descent for decentralized learning," *arXiv e-prints*, pp. 1–13, Aug. 2018.

[46] S. Sundaram and B. Gharesifard, "Consensus-based distributed optimization with malicious nodes," in *Proc. 53rd Annu. Allerton Conf.*, Monticello, IL, Sept. 2015, pp. 244–249.

[47] Z. Yang and W. U. Bajwa, "RD-SVM: A resilient distributed support vector machine," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, Shanghai, China, Mar. 2016, pp. 2444–2448.

[48] S. Kar and J. M. F. Moura, "Consensus+innovations distributed inference over networks," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 99–109, May 2013.

[49] NAAQS Table. [Online]. Available: https://www.epa.gov/criteria-air-pollutants/naaqs-table

Lu Zhou, Kuo-Hui Yeh, Gerhard Hancke,
Zhe Liu, and Chunhua Su

Security and Privacy for the Industrial Internet of Things

An overview of approaches to safeguarding endpoints

INTERNET OF THINGS—ISTOCKPHOTO.COM/IAREMENKO
CIRCUITS—IMAGE LICENSED BY INGRAM PUBLISHING

Endpoint devices form a core part of the architecture of the Industrial Internet of Things (IIoT). Aspects of endpoint device security also extend to related technology paradigms, such as cyberphysical systems (CPSs), edge computing, and fog computing. In this sphere, there have been several initiatives to define and promote safer and more secure IIoT networks, with the Industrial Internet Consortium (IIC) and OpenFog Consortium having developed security framework specifications detailing the techniques and technologies to secure industrial endpoints.

One of the core security mechanisms required for secure endpoints is cryptographic algorithms. Although there is a mature set of algorithms available, challenges remain in terms of efficient cryptographic algorithm implementation in the context of various constraints associated with the IIoT—not unlike the issues surrounding the efficient implementation of functions for digital signal processing. Endpoints are largely heterogeneous, with a wide range of overarching applications and resources, and therefore need efficient implementation approaches about logical processing, memory required, and execution time. This article aims to provide a high-level introduction to IIoT endpoint security requirements followed by a discussion on cryptographic algorithm implementation. Finally, we examine some system-wide design considerations for data security and privacy in current and emerging system designs.

Standardizing IIoT systems

The IIoT combines IoT technology with industrial CPSs (ICPSs), linking information and operational technology to offer improved system performance and data analytics [1]. ICPSs represent the integration of physical processes (e.g., actuation, control, and sensing) with communication and processing capabilities using interconnected embedded devices equipped with computational and communications technology [2], [3]. Industrial systems offer several technological challenges, setting safety, security, and resilience requirements for all of the components within the network architecture beyond those required in consumer technology sectors [2]–[4]. Standardization of IIoT systems, tied into accepted specifications for devices

Digital Object Identifier 10.1109/MSP.2018.2846297
Date of publication: 28 August 2018

and protocols within these systems, is essential for their deployment in industrial processes with strict performance and safety requirements. The usual size of IIoT deployments requires that future solutions developed for the IIoT should be highly scalable, with an extended operational period. The IIoT network should also be able to meet strict real-time deadlines, which means that communication latency and algorithm execution time should be predictable and minimal.

The challenges with regard to IIoT devices within ICPSs also extend to system security [5]. IIoT endpoint devices are an attractive target for attacks, and therefore it is critical that we protect the large-scale and often unmonitored deployment of devices [6]. Securing endpoint devices is made more difficult by a variety of device hardware and system restrictions, including limited device energy, memory, and processing resources; communication latencies; message size; and real-time operation [4]. Implementing traditional security techniques might fail, as the time the device dedicates to executing these techniques will delay the handling of its core industrial function, which could be unacceptable in time-critical industrial contexts.

Efficient security solutions for the IIoT (and other technology contexts) that are capable of securing systems regardless of limitations on power consumption, processing capacity, and memory footprint are of a high priority if we are to satisfy the security expectations that users and operators have for industrial applications. Security solutions for endpoint devices also apply to other expanding technologies in industrial applications, such as edge and fog computing. The fog computing paradigm aims to enable real-time analysis and faster actuation of sensor data by moving computation, control, and storage closer to the network edge in an IIoT network [7].

In this article, we focus on an overview of 1) efficient cryptography for IIoT endpoints, 2) scalable key management, and 3) system privacy issues. Networked endpoint devices, like programmable logic controllers, are an area of particular concern. These devices are vulnerable to physical tampering; a typical deployment often leaves devices unattended and a target for remote logical attacks, as they offer a stepping-stone to access the wider system. In addition, the size of IIoT deployments requires security mechanisms to be scalable. Thus, the first point of this article is endpoint device security. We introduce existing cryptographic mechanisms for the IIoT and discuss efficient algorithm implementation (resources and execution time). Traditional signal processing and IIoT cryptographic algorithm implementation display similarities, given the challenges of optimizing cryptography's underlying mathematical operations for resource efficiency and speed.

The second point is device key management. To facilitate the use of any security mechanisms, the system must manage, generate, and establish cryptographic keys to endpoint and intermediate gateway devices. However, keys are hard to establish and manage, even in small systems, with IIoT-required scalable solutions. In this area, we discuss key management approaches (public-key infrastructure for the IIoT and so forth), positives, negatives, and challenges to resolve, linking to the endpoint device security section with regard to realistic device needs and capabilities. Key management could be achieved in several ways, with increasing willingness to adopt public-key infrastructure (using certificates or not) if IIoT devices can be made to implement this efficiently.

The final IIoT issue considered here is privacy management. Privacy could relate to device-level (directly linked to a specific user) or system-level data (information about a specific user is inferred from multiple sources). Although the IIoT deals mostly with industrial control and automation, there is in some cases an overlap with the consumer area, e.g., in smart meters and building automation. In these cases, the perception of security by people interacting with the system could be crucial to system acceptance and deployment. In addition, we demonstrate a high-level discussion on cryptography for system-level data security and privacy (industrial data integrity only or also confidentiality), including a discussion on mechanisms to ensure that the IIoT adheres to privacy standards and legal compliance (for IIoT overlap with consumers, such as smart meters and smart homes).

Security requirements of IIoT endpoint devices

The value attached to security requirements is often subjective and application specific. As such, instead of promoting our own opinions, we use as a basis two well-known specifications of detailed security requirements for IIoT endpoint devices. The IIC comprises commercial and academic members collaborating on technical aspects related to the IIoT and tries to set standards for the design and operation of IIoT networks. The reference architecture proposed by the IIC aims to make the industrial Internet easily accessible through widely applicable, standards-based, open-architecture frameworks [1]. This allows for interoperable technologies that can be easily integrated, thereby expanding industrial Internet networks more quickly into new application areas.

The consortium has also developed a general security architecture to be used in conjunction with the reference framework [4]. Six interoperational building blocks, organized within three layers, form the functional basis of the security framework. The top layer comprises four foundations: 1) endpoint protection, 2) communication and connectivity protection, 3) security monitoring and analysis, and 4) security configuration management. A similar reference architecture has been developed by the OpenFog Consortium [8]. The security pillar defined in the architecture specifically discusses the following important attributes required of a fog device: privacy, anonymity, integrity, trust, attestation, verification, and measurement.

The general security specifications developed by the IIC and the OpenFog Consortium largely overlap, as shown in Table 1. Whereas the IIC provides more general guidelines as to what security services should be included and the objectives they should meet, the OpenFog Consortium provides more specific recommendations as to the technical mechanisms that could be used to provide a subset of these services. Looking at the two architectures in combination provides a good indication of what is expected of a secure IIoT endpoint device. What can be seen from both specifications is that cryptographic

Table 1. The IIC and OpenFog security objectives and recommendations [1], [4].

Functions	Security Objectives		Security Recommendations
	IIC	OpenFog	
Physical	X	X	Tamper resistance, evidence, detection, and response
Trust	X	X	Hardware root of trust, secure or verified boot, remote attestation, and secure boot processes
Identity	X	X	Credentials and immutable identifier with attestation
Access control	X	X	Authentication (cryptographic) and authorization
Integrity protection	X	X	Secure boot and run-time integrity checking and introspection
Data protection	X		Data confidentiality and integrity (cryptographic)
Monitoring and analysis	X		Detect anomaly events
Configuration and management	X		Signed software (cryptographic)
Cryptographic techniques	X	X	Symmetric encryption, message authentication and hash asymmetric encryption: integer and elliptic curve (ECDH, ECDA, ECQV) secure key generation and storage
Isolation techniques	X		Trusted execution environment/hypervisor

mechanisms are required and play an important role in several security functions, such as access control (authentication), device configuration and management, and data protection.

Any cryptographic solution will come at a cost, either in terms of additional device resources or system processing delay. IIoT devices are often highly resource constrained, in comparison to traditional information and communication technology equipment, and are required to operate at low power for months or years after their initial deployment. Although performance may improve with the use of new-generation IIoT processors [9], some cryptography implementations are unsuitable for use on legacy devices. For example, with software cryptographic algorithm implementations, large increases in memory occupation, execution time, and power consumption can be observed, particularly with older-generation devices. Similarly, adding cryptographic mechanisms has the potential to take up resources and introduce delays, thus making a device unable to operate in the real-time, mission-critical manner required.

With all cryptographic mechanisms, an appropriate implementation is needed to ensure that devices can provide security services while maintaining system functionality and ensuring that endpoint devices remain at a realistic cost point. Efficient cryptographic implementation in terms of execution time, resource cost, and energy consumption is therefore an important technical challenge that needs to be addressed for IIoT endpoint devices.

Cryptographic solutions for endpoint devices

Most IIoT endpoint devices are equipped with embedded processors that have limited computation resources and memory footprints. Endpoint equipment is often deployed in critical areas, meaning there is not only a need to communicate and authenticate with the control center but also between the devices themselves. On the other hand, an attacker may be able to access such equipment and perform various physical attacks, e.g., side-channel cryptanalysis. The integration of side-channel-resistant cryptographic solutions to secure the communication and computation inside the devices is a nontrivial task, because of the resource constraints and particularly the limited energy of endpoint devices.

Symmetric key cryptography refers to algorithms where the same secret cryptographic key is used for both encryption and decryption. There are three categories of symmetric cipher, depending on their concrete functions: 1) block cipher, 2) stream cipher, and 3) hash function. The idea behind a block cipher is to first partition the plaintext into relatively larger blocks—e.g., 128 bits for the Advanced Encryption Standard (AES)-128—and to further encode each of the blocks separately. A stream cipher, in contrast, is a symmetric key cipher where plaintext digits are combined with a pseudorandom cipher digit stream or keystream (e.g., RC4). A hash function is any function that can be used to map data of arbitrary size to data of fixed size (e.g., SHA3). The latter does not require a secret key, although it can be combined with a key to build symmetric cryptographic algorithms, such as HMAC.

From an algorithm standpoint, more than 20 lightweight ciphers have been designed and used in some industrial products since the 1990s. For example, the lightweight ciphers A5/1, A5/2, and ORYX designed in the 1990s have been used in cell phones, and the ciphers Hitag2 (designed in 2012) and Megamos (designed in 2013) have been adopted in car keys. For more information about lightweight symmetric ciphers for industry, see [10, Sec. 3.1].

From an implementation perspective, most of the lightweight ciphers have been designed with special implementation properties, e.g., Hight, Clefia, DESXL, and Present. An implementation of Hight requires approximately the same chip size as the AES algorithm (3,048 versus 3,400 gate equivalents), but the former is much faster. Most of the lightweight symmetric ciphers have the core structure of ARX-based and bitsliced-S-Box-based designs and simple key schedules, thus requiring less memory footprint while achieving fast execution time. For a comparison of hardware implementation among different lightweight symmetric ciphers, please refer to [11].

Microcontroller units are increasingly equipped with encryption hardware accelerators for standardized symmetric encryption and hash algorithms, such as AES, 3DES, SHA, and true random number generator (TRNG). However, asymmetric cryptography, often proposed for device and message authentication and key exchange, is still quite expensive when it comes to directly integrating it with IIoT endpoint devices. The next section focuses mainly on the lightweight implementation of elliptic curve cryptographic algorithms for IIoT endpoint equipment.

Asymmetric cryptographic algorithms for IIoT endpoint devices

Asymmetric cryptography, also commonly referred to as *public-key cryptography*, offers scalable solutions for key exchange and digital signatures, which are important in large IIoT networks. Key exchange can be seen as a method to securely establish the secrecy key via a public channel. The Diffie–Hellman (DH) key exchange, first published by Whitfield Diffie and Martin Hellman, is one of the earliest practical examples implemented in the field of cryptography. The security of the DH key exchange is based on the hardness of the discrete logarithm problem. RSA, first published by Rivest, Shamir, and Adleman, is based on the hardness of the integer factorization problem (IFP) and allows for encryption and digital signatures.

The key point of any software implementation of a public-key cryptographic scheme for endpoint devices is to find a suitable compromise between the following four requirements: 1) short execution time, 2) high flexibility and scalability (i.e., the support of curves providing different levels of security), 3) low memory—i.e., random access memory (RAM)—footprint, and 4) some basic protection against passive implementation attacks. Energy is, in general, the most precious resource of a battery-powered endpoint device. Compared to RSA and DH public-key cryptography, elliptic curve cryptography (ECC) is a lightweight public-key cryptography that was used by Neal Koblitz and Victor Miller in the 1980s. Its security is based on the intractability of the elliptic curve discrete logarithm problem (ECDLP), which allows one to use much smaller groups (compared to its classical counterpart, RSA based on the IFP). For example, it is generally accepted that ECC, instantiated with a 160-bit elliptic curve group, provides about the same level of security as the RSA signature scheme using a 1,024-bit modulus. Moreover, ECC has short-sized public/private key pairs and a smaller memory footprint. These features make ECC more suitable for use in the IIoT.

ECC can be used to implement key exchange and digital signatures more efficiently than classical DH and RSA. Elliptic curve DH (ECDH) is an anonymous key agreement protocol that allows two parties, each having an elliptic curve public/private key pair, to establish a shared secret over an insecure channel. The elliptic curve digital signature algorithm (ECDSA) offers a variant of the digital signature algorithm that uses ECC. The ECDSA can be used to provide entity and data-origin authentication, integrity protection, and nonrepudiation services, which makes it an essential tool for enabling secure communication. Common security protocols, such as AES, 3DES, and SHA, in addition to true random number generators often used in device-to-back end or gateway-to-back end communication within the IIoT, rely on these security algorithms to authenticate the server to the client (and, optionally, the client to the server) and to securely exchange the public keys needed for the establishment of an ephemeral shared secret.

Most of the current elliptic curve standards—e.g., the National Institute of Standards and Technology (NIST) curve and the IEEE P1363 curve—have adopted the form of a Weierstrass curve, and all of these standards rely on the fact that the ECDLP is difficult. However, the security of real-world ECC on IIoT devices does not only mean the security of the ECDLP but also the security of concrete implementations. For example, the widely adopted NIST P-256 curve is not considered to be a safe curve and fails to provide the features of complete point addition formulas and indistinguishability from uniform random strings. In the past ten years, researchers have paid a great deal of attention to evaluating new elliptic curve models. Some examples of well-studied curve models are the Montgomery model [12] and the twisted Edwards curve [13].

On the other hand, more than 15 years have passed since the standard curves were developed, and the cryptography community now has a better understanding of ECC security and practical implementation issues. The current state of the art has advanced. In research and other standards venues, newer variants of cryptographic schemes have been proposed that pursue better performance and/or simpler and more secure implementations. For example, MoTE-ECC is a novel approach for the implementation of ephemeral ECDH key exchange that exploits the birational equivalence of the Montgomery and twisted Edwards curves. By taking the individual computational advantages of the two forms into account, MoTE-ECC achieves higher performance (and better energy efficiency) and is also secure against basic side-channel attacks (e.g., timing attacks and simple analysis attacks). The Edwards-curve digital signature algorithm (EdDSA) is a state-of-the-art signature scheme using elliptic curves in twisted Edwards form that was developed with the intention of achieving both high performance (especially in software) and high security [14], [15]. A variant of the EdDSA is specified in RFC 8032 [16] and will be one of the signature algorithms supported in the next version of the TLS protocol, i.e., TLS 1.3.

From an arithmetic point of view, ephemeral ECDH key exchange between two sensor nodes requires each node to perform two scalar multiplications: one fixed-point scalar multiplication to generate an ephemeral key pair and another random-point scalar multiplication to obtain the shared secret. ECDSA requires one fixed-point scalar multiplication $k \cdot P$ to perform signature signing, while the verification process is relatively computation intensive, requiring a double scalar multiplication with a form of $k \cdot P + l \cdot Q$, where k and l are positive integers called *scalar* and P and Q are points on an elliptic curve E over a finite field F_p. Thus, efficient implementation of scalar multiplication is critical for cryptographic schemes.

As shown in Figure 1, an ECC implementation can be implemented in four layers: 1) cryptographic protocols (e.g., ECDH or ECDSA), 2) scalar multiplication, 3) group arithmetic (e.g., point doubling and point addition), and 4) field arithmetic (e.g., multiplication and addition). In the past 15 years, much research has been done to improve the performance of

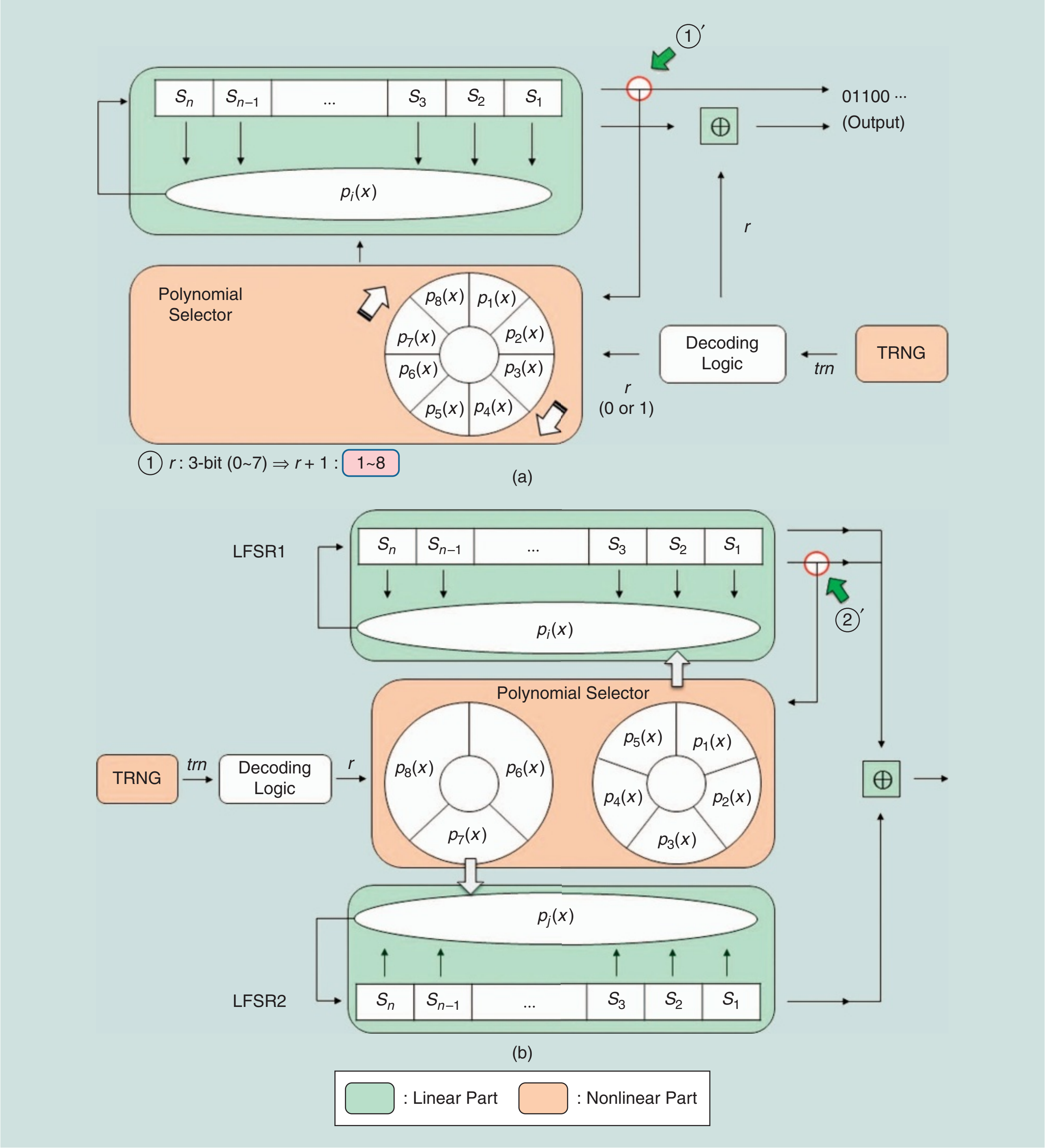


FIGURE 1. The lightweight key generation for IoT devices in (a) proposal 1 and (b) proposal 2. (Figure courtesy of [24].)

elliptic curve operations on 8- and 16-bit microcontrollers, making ECC more attractive for resource-constrained environments. Most of the work improved the performance of scalar multiplication either by proposing the performance of field arithmetic (e.g., field multiplication) or by choosing the special family of the underlying fields or elliptic curve models.

The first research line is to propose new variants of multiprecision arithmetic and focus on improving the standardized elliptic curve. The first really efficient ECC software for an 8-bit microcontroller was introduced in [17] at the Cryptographic Hardware and Embedded Systems Conference in 2004. In their work, Gura et al. introduced the first optimized multiprecision multiplication for small embedded devices, which they called *hybrid multiplication*. This combines the advantages of both the operand- and product-scanning methods and was the first multiprecision platform-specified arithmetic that carefully optimized the number of addition-with-carry and memory-access instructions. Based on this classic method, the researchers reported an execution time of only $6.48 \cdot 10^6$ clock cycles for a full scalar multiplication over a 160-bit SECG-compliant prime field on IIoT endpoint devices.

In the 14 years since the publication of their seminal paper, a large body of research has been devoted to further reducing the execution time of ECC on IIoT devices. The majority of this work has focused on advancing the multiprecision arithmetic operation or devising more efficient variants of it. For example, Lederer et al. [18], presenting at the 2009 Workshop in Information Security Theory and Practice, improved Gura et al.'s work to further reduce the number of addition-with-carry instructions by reorganizing the byte multiplication in the inner loop and then implementing ECDH key exchange using a NIST P-192 curve. Their implementation requires an execution time of $12.33 \cdot 10^6$ cycles for a random base point scalar multiplication and $5.20 \cdot 10^6$ cycles when the base point is fixed and known a priori.

Besides implementation on 8-bit endpoint devices, another platform that frequently sees endpoint devices used is the MSP430 series of microcontrollers produced by Texas Instruments. On such 16-bit platforms, Wang et al. [31] reported one of the first ECC software implementations on a Weierstrass curve defined over a 160-bit prime field, in which the execution time was 25 and 28.1 million cycles for a fixed-base and a variable-base scalar multiplication, respectively. Some well-known libraries on endpoint devices are TinyECC, WM-ECC, and Nano-ECC, all of which are highly scalable and configurable and support Weierstrass curves defined over 128-, 160-, and 192-bit prime fields.

Another research line is to employ a special family of prime fields or elliptic curves to further reduce the energy consumption of the elliptic curve key exchange and signature. One classic ECC software implementation for an endpoint device equipped with an 8-bit microcontroller was reported by Woodbury et al. in 2000. The authors chose a special family of fields called *optimal extension fields* (*OEFs*), which are finite fields each consisting of p^m elements, where p is a pseudo-Mersenne prime (i.e., a prime of the form $p = 2^k - c$) and m is chosen such that an irreducible binomial $x(t) = t^m - \omega$ exists over GF(p). The specific OEF is $\mathrm{GF}((2^8 - 17)^{17})$, which allows the arithmetic operations, especially the multiplication and inversion, to be executed efficiently with small devices. Their implementation requires an execution time of roughly $100 \cdot 10^6$ clock cycles for random-point scalar multiplication.

Liu et al., during the 2013 International Conference on Information Security and Cryptology, adopted optimal prime fields (OPFs) and studied the suitability of OPFs for a lightweight implementation of ECC with a view toward high performance and security. The researchers proposed a performance-optimized implementation using a Montgomery curve and a security-optimized implementation with a GLV curve on an 8-bit IIoT platform. Later, in [19], Liu et al. presented the design of a scalable, regular, and highly optimized ECC library using a MoTE curve for both MICAz and Tmote Sky IIoT endpoint devices, which supports widely used key-exchange and signature schemes. Their parameterized implementation of elliptic curve group arithmetic supports pseudo-Mersenne prime fields at different security levels with two optimized-specific designs: the high-speed version and the memory-efficient version.

Some other well-known fast and secure ECC implementations on endpoint devices include field-programmable gate array implementation of signature verification operation [20], the NaCl library, Curve25519 implementation [21], and the recently proposed FourQ [22] and memory-efficient ECC [23] libraries. We summarize the execution times of existing implementations in Table 2.

Device key management for the IIoT

IIoT devices are commonly used in and facilitate the application of various wireless communication technologies for small devices with low-cost hardware and software interfaces. At the same time, secure communication and related applications rely on the security of key management inside the IoT devices and their supporting environment. An attacker always wants to compromise a device and get the secret key via communication interception, side-channel analysis, or reverse engineering. If the attacker can

Table 2. The execution times of existing ECC-based implementations for IIoT endpoint devices.

Method	Execution Time
Gura et al. [17], 2004	$6.48 \cdot 10^6$ clock cycles on 8-bit AVR microcontroller (80-bit security level)
Lederer et al. [18], 2009	$5.20 \cdot 10^6$ on 8-bit AVR microcontroller (96-bit security level)
Liu et al. [19], 2017	$8.59 \cdot 10^6$ clock cycles on 8-bit AVR processors (software implementation and 128-bit security level), and $6.10 \cdot 10^6$ clock cycles on 16-bit MSP430 processors (software implementation and 128-bit security level)
Liu et al. [20], 2017	$1.8 \cdot 10^6$ clock cycles (hardware–software codesign and 102-bit security level)
Düll et al. [21], 2015 (fixed-point scalar multiplication)	$7.0 \cdot 10^6$ clock cycles (software implementation, on 8-bit AVR processor and 128-bit security level), $4.5 \cdot 10^6$ clock cycles (software implementation on 16-bit MSP430 processor and 128-bit security level), and $1.8 \cdot 10^6$ clock cycles (software implementation on 32-bit ARM processor and 128-bit security level)
Liu et al. [22], 2017 (fixed-point scalar multiplication)	$2.9 \cdot 10^6$ clock cycles (software implementation on 8-bit AVR processor and 128-bit security level), $1.8 \cdot 10^6$ clock cycles (software implementation on 16-bit MSP430 processor and 128-bit security level), and $0.23 \cdot 10^6$ clock cycles (software implementation on 32-bit ARM processor and 128-bit security level)
Liu et al. [23], 2018	$1.6 \cdot 10^6$ clock cycles on 32-bit ARMv6-M processor (128-bit security level)

manage to reveal the device key, the time needed is substantially reduced. In such an event, if a mechanism exists to deactivate this compromised key, the potential risk from the attack could be mitigated. In a malicious IoT application environment, we need to establish secure device key management technology to defend our IoT-enabled industrial applications.

The key in each IIoT device has its individual lifetime cycle, and the key management involves managing various key lifetime cycles for many IIoT devices. The lifetime cycle includes the random key bit generation, key distribution among devices, key storage, and key update and revocation. IIoT devices' secure key management is quite challenging and extraordinarily difficult to implement when there are a great many unattended devices. There are some challenging issues we must clarify, as follows.

- The devices are produced by different external manufacturers, so it is necessary for them to be provisioned with cryptographic keys, and those keys must be protected once provisioned. Different key sizes provide different security levels.
- IoT devices can be more easily hacked compared to conventional computing devices (such as a personal computer) or tamperproof devices (such as a smart card), so the update mechanism should be robust and capable of providing key recovery functions.
- IoT devices are resource restrained, and, for this reason, it is difficult to employ conventional cryptography-based key management schemes directly on IoT devices.

Cryptography is one of the fundamental primitives in IIoT secure key management. Cryptographic techniques are applied after the keying material is agreed upon in advance in the communicating IIoT devices. As the main task of the IIoT key management protocols, the key management mechanism is either centralized, decentralized, or distributed for IoT applications. Centralized solutions are based on a centralized implementation called a *key distribution center* (*KDC*), which produces and distributes the keys to all of the IoT devices. Decentralized solutions operate on a network partitioned into a fixed number of small groups where each group has a managing device. The functionality of the KDC is to share the keys between the group-managing devices, which are usually organized in a hierarchical structure. For distributed solutions, nodes collaborate to ensure the key management operations, such as key generation, distribution, renewal, and revocation.

Secure key generation

The secret keys in many IIoT devices must be preinstalled. However, this method is vulnerable to adversaries who can reverse-engineer hardware or software to obtain the secret keys, so it is preferable for the keys to be updated by the devices themselves. In such cases, for the general purpose of lightweight key generation for the IIoT, we should design mechanisms that satisfy the following properties.

- *Low resource consumption*: The resource consumption of both hardware and software for generating the pseudorandom key bits must be low because of the limited power available to IIoT devices.
- *Low memory requirement*: The amount of information stored in IIoT devices should be kept as small as possible since the equipment's memory is normally extremely constrained.

Many existing solutions are based on lightweight cryptography and utilize linear feedback shift register (LFSR) designs to keep the cost low. To give an overview of these designs, we discuss as examples two proposals for key generation in lightweight IIoT devices [24]. The first is based on modified multiple LFSRs' pseudorandom number generators. The basic idea is to make a random choice from eight 16-bit LFSRs. It is inspired by Sugei's J3Gen scheme [25], where the feedback polynomials are implemented as a wheel that rotates depending on the bit value given by the TRNG module. If the truly random bit is a logical 0, the wheel rotates one position; that is, it selects the next feedback polynomial. Conversely, if the truly random bit is a logical 1, then the wheel rotates two positions; that is, the polynomial selector jumps over one feedback polynomial and selects the next one. The first proposal modifies this, as shown in Figure 1(a).

In the second case, shown in Figure 1(b), the randomized key bits are loaded into two independent registers, and the randomization is executed there. Our proposal is inspired by a well-known lightweight stream cipher, KTANTAN [26], and the random key bits are derived from the LFSR while doing randomization. For each round, some bits are taken from the LFSR and input in the mixing process, or two nonlinear Boolean functions can be used. Our simple construction is modified by adding internal random bits instead of using the computationally costly nonlinear functions. The Boolean functions used in our construction will output random bits, which are loaded to the least significant bits of the registers, after the internal key bits are shifted. This should be done in an invertible manner. To ensure sufficient randomization when generating the key bits, the devices should wait for several rounds of the LFSR processing to be executed. After that, the devices can obtain key bits with higher security.

In many IIoT applications, the input seed of the internal pseudorandom number generator (PRNG) state is loaded once and fixed inside the device, which is a vulnerability allowing adversaries to obtain it to predict the key bits. To improve the security in such cases and avoid the attack, our proposal provides an efficient randomized approach that makes the input seed not be stored in memory. We also construct new internal operations for XOR expressions for the irreducible polynomials used in our PRNG. As a routine in utilizing security primitives for IIoT devices, many solutions use the XOR operator, whose implementation is inexpensive, usually using only some LFSR.

The degree of an XOR expression depends on the number of distinctly named variables in an expression. We can observe that the sum of three irreducible polynomial expressions $x \oplus y \oplus z$ has a degree of three, but the sum remains linear, which requires nonlinear processing to make it more secure. The purpose of our proposal is to increase the degree for randomizing the internal state of key bits without increasing hardware and software resource consumption during the implementation. For some distinct bitwise variables inside the IIoT devices, we can select our customized reduced polynomial form if it is expressed as the minimum degree that still makes security analysis simple. We divide our chosen LFSR into two parts to lower the cost of implementation. At the same time, we attain a stronger security condition by doubling the internal state of the LFSRs with two combined 8-bit LFSRs, allowing the key bits generation to attain the full randomness of a 16-bit LFSR.

The reason for our irreducible polynomial assignment is to achieve efficient hardware implementation by choosing polynomials with several coefficients in common, with the common coefficients x_{16}, x_{11}, x_6, x_5, and x_0 shared by two irreducible polynomials. Our method simplifies the hardware construction with fewer gates. Furthermore, our method makes selection of these feedback polynomials more flexible and without potential key bit leakage. Using our method, we can employ a dynamic key server to also generate a key for an identity pattern without storing that key. Access to a key works the same way as with a static key server, except the key is generated again for subsequent retrieval. A dynamic key server depends on a functional derivation per the IIoT identity for a key. If the same identity is presented multiple times, the same key bits will be XORed with random bits.

Secure key storage and retrieval

Many IIoT devices are not tamper resistant and do not have access to trusted hardware modules. To protect the keys in such IIoT equipment, it is very important to hide the memory-access pattern in the devices themselves because adversaries can observe the memory read and write operations and get the key via a side-channel attack or Trojan virus. One possible approach is that whenever the IIoT device reads or makes an update for the key in memory, we make all of the key bits access pattern randomized when communicating with cloud servers or other IIoT devices. This can be accomplished with oblivious RAM (ORAM) schemes, as shown, for example, in the method in Figure 2 [27].

The accessed location of one key bit can be important information required by the IIoT device because, during the encryption and decryption processes, the key bits are the most accessed compared to normal data access. For this reason, the uploaded data blocks and the memory locations that contain the secret key bits should be different than what was previously downloaded. Whenever the IIoT device wants to access (read or write) key bits data, the address (x_i, y_j) is obtained from the position map. The device then reads H blocks from the

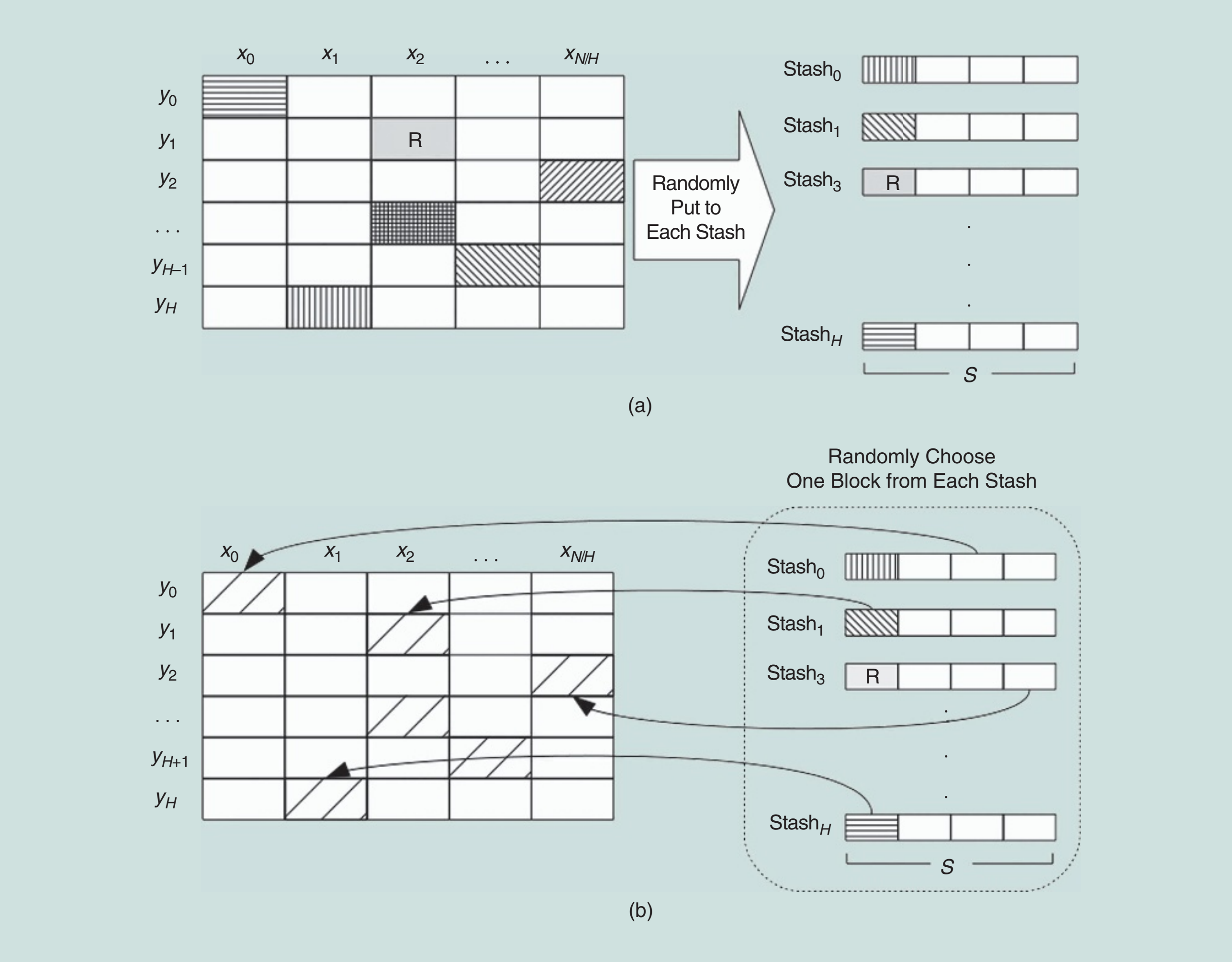


FIGURE 2. An illustration of an M-ORAM mechanism. (Figure courtesy of [27].)

server, one block from each row in the matrix. The devices will choose the memory units or data block via the columns and rows of the matrix, such as when the row is y_j, then the column is x_i. Otherwise, the data block locations are chosen uniformly in a random manner from the set of memory units accessed by the previous operation, and the columns are chosen in a uniformly random manner for the remaining rows.

The purpose of choosing columns more randomly (in addition to the block that includes the data of interest) is to ensure that adversaries cannot track the key bits' access pattern that contains the location of the key bits' storage in the memory. In our scheme, we also make some key bits' locations remain the same as with the previous memory access. In such cases, the adversary cannot distinguish whether previous and later access is different or not. That is, if we do not choose some addresses from the previous memory locations that access the same key bits, then accessing two different key bits would result in two different memory locations, allowing adversaries to identify the access patterns.

Compared to other existing ORAM schemes, we improve the security by making reencryption using AES after each time an IoT memory unit is accessed. After a memory unit with data block is downloaded, it will be decrypted in local IIoT devices. For the next data access, the devices use a new key. Therefore, the adversary cannot identify that the uploaded data are the same as those previously accessed. In matrix-based ORAM (M-ORAM), we can apply any encryption schemes where the data block and its identity are encrypted using a key generated from a pseudorandom function (PRF). Importantly, the PRF takes the data identity (unique to each memory unit), a common secret key for all memory units, and a counter that is associated with each memory unit as input, which reduces the resource occupation when making the key storage and access in IIoT devices.

Other issues: Privacy for the IIoT

The rapid advancement in wireless communications and the pervasive computing abilities of smart objects have brought about a new era of application development, from industrial control systems to critical IIoT infrastructure, providing intelligence and optimization of industry-related processes about resource utilization. With an increasing number of IIoT objects being equipped with technology to provide identification, computation, and communication capabilities during industrial operation processes, we also need to consider system-wide security and privacy issues. A secure endpoint should be part of an overall security approach that ensures data security for any interactions among endpoint devices (or smart things) and the back-end ICPSs, as shown in Figure 3. The privacy implications of system data, especially data originating with customers, should also be considered with respect to processing, storage, and access by system operators. When considering data security and privacy in the IIoT, there are three levels of sensitive (or private) data and information involved.

1) *User level*: This level involves access control to allow authorized persons to access appropriate-level data stored on cloud clusters (or objects). Examples of such data include real-time monitoring data and meaningful analyzed information. Useful protection techniques include identity-card-based login mechanisms and biometric authentication. In addition, a log corresponding to each access activity must be maintained for audit.
2) *Machine level*: At this level, data are stored and transmitted among multiple objects (and gateways) in an IIoT system. Snooping on the network to probe organization-oriented private data, such as identification and access history, is highly possible when data are transmitted across networks in an unprotected way. It is suggested to implement secure machine-to-machine (M2M) communications, device management, and automatic firmware updating to maintain data confidentiality and system robustness.
3) *System level*: An ICPS, consisting of physical and software components, acts as a computing platform that monitors and controls physical processes. The data processing in an ICPS is critical for the IIoT. This level involves data collected from machines and analyzed by the ICPS itself.

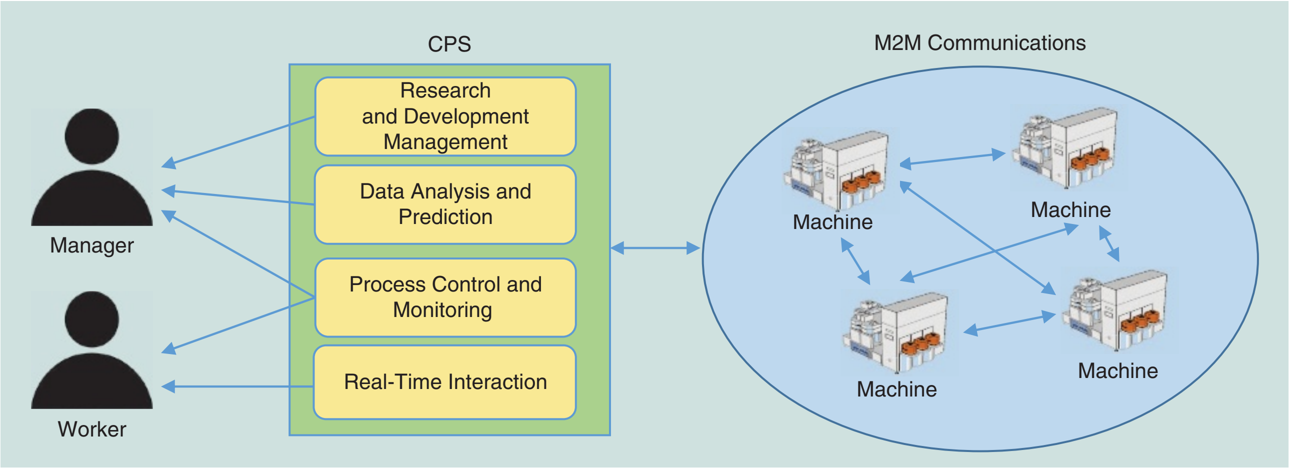

CPS
M2M Communications
Research and Development Management
Data Analysis and Prediction
Process Control and Monitoring
Real-Time Interaction
Manager
Worker
Machine
Machine
Machine
Machine

FIGURE 3. An illustration of the back end and human interface of the IIoT.

Without appropriate security mechanisms, organizational privacy leakage is unavoidable. Enhanced efforts on security architectures for CPSs are strongly suggested.

While the IIoT promises new opportunities for innovative service applications and business models through the effective use of next-generation mobile devices, it brings with it many challenges with respect to ICPSs, such as the Slammer worm, Stuxnet, and DUQU, as well as with regard to the endpoint (e.g., device and user), such as individual (or organizational) privacy concerns, social engineering, man-in-the-middle attacks, denial-of-service attacks, reverse engineering, malware, and side-channel attacks [6]. As a result, in terms of the enhancement of security and privacy for the IIoT, significant efforts have been dedicated to eliminating these potential vulnerabilities and threats. In the following section, we will discuss possible solutions, based on the aspects of data confidentiality, integrity, and authenticity; privacy protection on cloud servers with big-data analysis; and privacy management on end objects.

Data confidentiality, integrity, and authenticity

Cryptography techniques, such as encryption, hash functions, and digital signatures, constitute an important area when it comes to ensuring data confidentiality, integrity, and authenticity. However, computing resource limitations and the heterogeneity of IIoT objects give rise to critical new challenges, making it inevitable to reengineer traditional security mechanisms or even create new solutions to fit the specific requirements of the IIoT.

First, authenticated encryption is one of the most promising techniques to secure IIoT endpoint devices, as it is able to provide both confidentiality and authenticity of data while achieving high efficiency of computation and end-to-end communication. Recently, a process called *Competition for Authenticated Encryption: Security, Applicability, and Robustness* (*CAESAR*) was launched to search for a new authenticated encryption algorithm that can offer advantages over AES-GCM and is suitable for widespread use. So far, among the candidates in CAESAR's Round 3, the computational efficiency of Deoxys, which adopts tweakable block and linear transformations, has shown itself to be suitably efficient for implementation in IIoT applications. At the same time, the security density Deoxys provides is acceptable. Another candidate, called *CLOC & SILC*, is secure against partial nonce misuse and can provide an acceptable security level. In particular, excellent performance, i.e., computational efficiency and memory utilization, can be achieved with small-size data, and thus CLOC & SILC are suitable for M2M authentication.

What's more, with the rapid growth and universality of wearable devices, it is feasible to implement a continuous authentication scheme for an IIoT-based environment with users possessing wearables. New types of continuous authentication mechanisms, e.g., brain waves [28], have been realized to support continuous (or real-time) entity verification in the background without the need for direct input from the user. This shifts the retrieval of physical signals and biofactors for entity verification and authentication closer to the consumer end.

Privacy protection on cloud servers with big-data analysis

The use of predictive analytics to make useful decisions about individuals may have negative impacts. An illustrative example would be a case where, because of automated decision making, a company promoted baby-related products to an expectant mother before she announced to her family she was pregnant. Similar situations could arise regarding sensitive personal information about one's sexual orientation or health status. Moreover, it is increasingly difficult for organizations to anonymize data and simultaneously use them for individual identification. Hence, it is critical to protect individuals' privacy during data processing, and the following tenets are recommended as a baseline for privacy protection.

- Data must be processed fairly and used for specified and lawful purposes.
- Unauthorized or unlawful processing of data must be efficiently detected and dealt with.
- Accountability should be guaranteed.
- The consent obtained for data processing should be freely given.
- Data must not be exploited without an adequate level of protection.
- Data must be adequate, relevant, and not excessive in relation to the purpose for which they are processed.
- Data processed for any purpose must not be kept for longer than is necessary for that purpose.

Privacy management on endpoint devices

Data privacy on IIoT objects requires an effective control scheme to govern access to data stored inside these objects. It is recommended to extend traditional access control approaches to fine-grained, context-based access control systems in which IIoT objects can be dynamically controlled in terms of acquiring data based on context. In addition, implementing secure M2M communications among IIoT objects for data confidentiality is suggested. On the other hand, heterogeneous communication architectures are common in IIoT-oriented environments because various types of smart objects and relevant communication techniques, such as radio frequency, Bluetooth Low Energy (BLE), Zigbee, LoRa, and Wi-Fi, are adopted.

It is necessary to consider the robustness of the privacy protection schemes of IIoT-based communication techniques, such as BLE's random address technique and anonymous communication. Moreover, it is highly recommended to adopt standards for privacy protection in the IIoT. As far as the European Union (EU) is concerned, the future evolution of EU laws and directives regarding privacy and personal data protection will see a move toward a privacy-by-design legal framework [29], [30], where seven major processes are recommended.

1) *Proactive and preventative*: Anticipate and prevent privacy-invasive events before they happen.
2) *Privacy as the default*: Ensure the maximum degree of data protection and privacy preservation in the IIoT.

3) *Privacy embedded into design*: Privacy protection must be an essential component of the core IIoT system.
4) *Full functionality*: Preserving privacy must be accomplished without making any nonrelevant tradeoffs with security.
5) *End-to-end security*: All data relevant to the IIoT must be securely collected, retained, and destroyed at the end of the process, which represents the concept of secure life-cycle management of information.
6) *Visibility and transparency*: The user should know who possesses his/her data, what data have been collected and processed, and for what purposes.
7) *Respect for user privacy*: Offer users strong privacy defaults and appropriate notices with user-friendly options.

In addition, a complete process consisting of identification, preservation, collection, processing, review, analysis, and production for the management of electronically stored data is required to support auditing throughout the data life cycle.

Finally, wearable devices undeniably represent one of the most promising paradigms in terms of ubiquitous computing in IoT-enabled scenarios. Good examples include fitness bands (i.e., activity trackers), running watches, and wearable glasses that are capable of Internet connectivity, enabling the exchange of data without human intervention. In IIoT scenarios, individuals may be embedded with their own wearables during working periods. Therefore, it is necessary to take stock of the efficiency, attendant benefits, and security risks of so-called wear-your-own-device (WYOD) scenarios and to implement a WYOD model for management.

Conclusions

There are several initiatives for specifying security specifications and requirements for IIoT endpoints. One of the core security mechanisms required for secure endpoints is cryptographic algorithms. Although there is a mature set of algorithms available, challenges remain in terms of efficient algorithm implementation and associated key management in the context of the various constraints associated with the IIoT.

We presented a brief discussion on symmetric and asymmetric cryptographic algorithms. With the former increasingly being integrated using efficient cryptographic coprocessors, future research challenges lie more with the latter, which are still often implemented in device software, where they must compete for resources with other system processes. In this regard, ECC is a promising approach to providing both scalable key-exchange and digital signature mechanisms in large IIoT systems, and we provided an overview of current implementation approaches. Key management is also challenging on devices with limited resources and little to no trusted hardware. New methods for allowing devices to generate keys, in addition to storing and accessing them securely, are needed, and we provided examples of common lightweight approaches to LFSR-based key generation and oblivious random-access mechanisms. Finally, we concluded with a system-wide overview of data security and privacy issues that need to be considered in the IIoT, including future security issues related to big-data analysis and storage and data legal frameworks.

Acknowledgments

This work was supported by the Natural Science Foundation of Jiangsu Province of China (number SBK2018043466), the Fundamental Research Funds for the Central Universities (grant NE2018106), the Academia Sinica, Taiwan (grant MOST 106-3114-E-011-003), the Ministry of Science and Technology, Taiwan (grants MOST 105-2221-E-259-014-MY3 and MOST 105-2923-E-182-001-MY3), the City University of Hong Kong (project 9678131), and the Japan Society for the Promotion of Science KAKENHI (grant Kiban C 18K11298).

Authors

Lu Zhou (sduzhoulu@gmail.com) received her B.S. and M.S. degrees in computer science from Shandong University, China, with first-class honors in 2012 and 2015, respectively. She is a Ph.D. student in the Division of Computer Science, University of Aizu, Aizuwakamatsu, Japan. Her research interests include security privacy and cryptographic solutions for the Internet of Things, cloud computing security, and game theory. She has published more than ten peer-reviewed research papers in related major journals and conferences.

Kuo-Hui Yeh (khyeh@gms.ndhu.edu.tw) received his B.S. degree in mathematics from the Fu Jen Catholic University, New Taipei City, Taiwan, in 2000 and his M.S. and Ph.D. degrees in information management from the National Taiwan University of Science and Technology, Taipei, in 2005 and 2010, respectively. He is an associate professor with the Department of Information Management, National Dong Hwa University, Hualien, Taiwan. His research interests include Internet of Things security, blockchain, mobile security, near-field communication/radio-frequency identification security, authentication, digital signature, data privacy, and network security. He has authored more than 90 articles in international journals and conference proceedings. He is currently an associate editor of *IEEE Access* and *Journal of Internet Technology* (*JIT*) and has served as a guest editor of *Elsevier Future Generation Computer Systems*, *JIT*, *Sensors*, and *Cryptography*. In addition, he has served as a Technical Program Committee member of 24 international conferences and workshops on information security. He is a Senior Member of the IEEE.

Gerhard Hancke (ghancke@ieee.org) received his B.Eng. and M.Eng. degrees in computer engineering from the University of Pretoria, South Africa, in 2002 and 2003, respectively, and a Ph.D. degree in computer science from the Computer Laboratory, University of Cambridge, United Kingdom, in 2008. He worked at the Information Security Group, University of London, where he was involved in the evaluation, development, and integration of smart card and radio-frequency identification systems. He has been an assistant professor with the Department of Computer Science, City University of Hong Kong, China, since 2013. He is chair of the IEEE Industrial Electronics Society's Technical Committee on Cloud and Wireless Systems for Industrial Applications and is an associate editor of *IEEE Transactions on Industrial Informatics*. He has written approximately 100 academic papers on system security and industrial applications for the Internet of Things and cyberphysical systems. He is a Senior Member of the IEEE.

Zhe Liu (sduliuzhe@gmail.com) received his B.S. and M.S. degrees in computer science from Shandong University, China, in 2008 and 2011, respectively. He received his Ph.D. degree in applied cryptography from the University of Luxembourg, in 2015, and the prestigious FNR Outstanding Ph.D. Thesis Award in 2016. He is a professor in the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China, and a researcher at the Interdisciplinary Centre for Security, Reliability, and Trust, University of Luxembourg. His research interests are cryptographic engineering and security of the Internet of Things. He has published more than 70 peer-reviewed research papers.

Chunhua Su (chsu@u-aizu.ac.jp) received his B.S. degree from the Beijing Electronic and Science Institute in 2003 and his M.S. and Ph.D. degrees in computer science from the Faculty of Engineering, Kyushu University, Japan, in 2006 and 2009, respectively. He worked with the Institute for Infocomm Research, Singapore, from 2011 to 2013. From 2013 to 2017, he was an assistant professor at the Japan Advanced Institute of Science and Technology, Osaka University. He is currently an associate professor in the Division of Computer Science, University of Aizu, Aizuwakamatsu, Japan. His research interests include cryptanalysis, cryptographic protocols, privacy-preserving technologies in data mining, and Internet of Things security and privacy.

References

[1] Industrial Internet Consortium. (2015, June 4). Industrial Internet reference architecture. [Online]. Available: http://www.iiconsortium.org/IIRA-1-7-ajs.pdf

[2] M. Wollschlaeger, T. Sauter, and J. Jasperneite, "The future of industrial communication: Automation networks in the era of the Internet of Things and industry 4.0," *IEEE Ind. Electron. Mag.*, vol. 11, no. 1, pp. 17–27, 2017.

[3] B. Cheng, J. Zhang, G. P. Hancke, S. Karnouskos, and A. W. Colombo, "Industrial cyberphysical systems: Realizing cloud-based big data infrastructures," *IEEE Ind. Electron. Mag.*, vol. 12, no. 1, pp. 25–35, 2018.

[4] Industrial Internet Consortium. (2016). Industrial Internet of Things volume G4: Security framework. [Online]. Available: http://www.iiconsortium.org/pdf/IIC_PUB_G4_V1.00_PB-3.pdf

[5] D. Gollmann and M. Krotofil, "Cyber-physical systems security," in *The New Codebreakers*, P. Ryan, D. Naccache, and J.-J. Quisquater, Eds. Berlin: Springer-Verlag, 2016, pp. 195–204.

[6] A. R. Sadeghi, C. Wachsmann, and M. Waidner, "Security and privacy challenges in industrial Internet of Things," in *Proc. 52nd ACM/EDAC/IEEE Design Automation Conf. (DAC)*, 2015, pp. 1–6.

[7] B. Tang, Z. Chen, G. Hefferman, S. Pei, T. Wei, H. He, and Q. Yang, "Incorporating intelligence in fog computing for big data analysis in smart cities," *IEEE Trans. Ind. Informat.*, vol. 13, no. 5, pp. 2140–2150, 2017.

[8] OpenFog Consortium Architecture Working Group, "OpenFog reference architecture for fog computing," OpenFog Consortium, Fremont, CA, Tech. Rep. OPFRA001.020817, 2017.

[9] L. P. I. Ledwaba, G. P. Hancke, H. S. Venter, and S. J. Isaac, "Performance costs of software cryptography in securing new-generation Internet of energy endpoint devices," *IEEE Access*, vol. 6, pp. 9303–9323, Jan. 2018.

[10] A. Biryukov and L. P. Perrin. (2017). State of the art in lightweight symmetric cryptography. [Online]. Available: http://orbilu.uni.lu/handle/10993/31319

[11] T. Eisenbarth and S. Kumar, "A survey of lightweight-cryptography implementations," *IEEE Des. Test. Comput.*, vol. 24, no. 6, pp. 522–533, 2007.

[12] P. L. Montgomery, "Speeding the Pollard and elliptic curve methods of factorization," *Math. Comput.*, vol. 48, no. 177, pp. 243–264, 1987.

[13] D. J. Bernstein, P. Birkner, M. Joye, T. Lange, and C. Peters, "Twisted Edwards curves," in *Progress in Cryptology—AFRICACRYPT 2008* (Lecture Notes in Computer Science Series), vol. 5023, S. Vaudenay, Ed. Berlin: Springer-Verlag, 2008, pp. 389–405.

[14] D. J. Bernstein, N. Duif, T. Lange, P. Schwabe, and B.-Y. Yang, "High-speed high-security signatures," in *Cryptographic Hardware and Embedded Systems—CHES 2011* (Lecture Notes in Computer Science), vol. 6917, B. Preneel and T. Takagi, Eds. Berlin: Springer-Verlag, 2011, pp. 124–142.

[15] D. J. Bernstein, N. Duif, T. Lange, P. Schwabe, and B.-Y. Yang, "High-speed high-security signatures," *J. Cryptog. Eng.*, vol. 2, no. 2, pp. 77–89, 2012.

[16] S. Josefsson and I. Liusvaara, "Edwards-curve digital signature algorithm (EdDSA)," Internet Research Task Force, Crypto Forum Research Group, Tech. Rep. RFC 8032, 2017.

[17] N. Gura, A. Patel, A. S. Wander, H. Eberle, and S. Chang Shantz, "Comparing elliptic curve cryptography and RSA on 8-bit CPUs," in *Cryptographic Hardware and Embedded Systems—CHES 2004* (Lecture Notes in Computer Science Series), vol. 3156, M. Joye and J.-J. Quisquater, Eds. Berlin: Springer-Verlag, 2004, pp. 119–132.

[18] C. Lederer, R. Mader, M. Koschuch, J. Großschädl, A. Szekely, and S. Tillich, "Energy-efficient implementation of ECDH key exchange for wireless sensor networks," in *Information Security Theory and Practice—WISTP 2009* (Lecture Notes in Computer Science Series), vol. 5746, O. Markowitch, A. Bilas, J.-H. Hoepman, C. J. Mitchell, and J.-J. Quisquater, Eds. Berlin: Springer-Verlag, 2009, pp. 112–127.

[19] Z. Liu, X. Huang, Z. Hu, M. K. Khan, H. Seo, and L. Zhou, "On emerging family of elliptic curves to secure Internet of Things: ECC comes of age," *IEEE Trans. Depend. Secure Comput.*, vol. 14, no. 3, pp. 237–248, 2017.

[20] Z. Liu, J. Großschädl, Z. Hu, K. Järvinen, H. Wang, and I. Verbauwhede, "Elliptic curve cryptography with efficiently computable endomorphisms and its hardware implementations for the Internet of Things," *IEEE Trans. Comput.*, vol. 66, no. 5, pp. 773–785, 2017.

[21] M. Düll, B. Haase, G. Hinterwälder, M. Hutter, C. Paar, A. H. Sánchez, and P. Schwabe, "High-speed Curve25519 on 8-bit, 16-bit and 32-bit microcontrollers," *Designs, Codes Cryptog.*, vol. 77, no. 2–3, pp. 493–514, 2015.

[22] Z. Liu, P. Longa, G. C. C. F. Pereira, O. Reparaz, and H. Seo, "FourQ on embedded devices with strong countermeasures against side-channel attacks," in *Proc. 19th Int. Conf. Cryptographic Hardware and Embedded Systems*, 2017, pp. 665–686.

[23] Z. Liu, H. Seo, A. Castiglione, K.-K. R. Choo, and H. Kim, "Memory-efficient implementation of elliptic curve cryptography for the Internet-of-Things," *IEEE Trans. Depend. Secure Comput.*, 2018, to be published.

[24] J. Chen, A. Miyaj, H. Sato, and C. Su, "Improved lightweight pseudo-random number generators for the low-cost RFID tags," in *Proc. 14th IEEE Int. Conf. Trust, Security and Privacy Computing and Communications (Trustcom),* 2015, vol. 1, pp. 17–24.

[25] J. Melia-Segui, J. Garcia-Alfaro, and J. Herrera-Joancomarti. (2013). J3Gen: A PRNG for low-cost passive RFID. *Sensors*. [Online]. *13(3)*, pp. 3816–3830. Available: http://www.mdpi.com/1424-8220/13/3/3816

[26] C. De Cannire, O. Dunkelman, and M. Kneevi. (2009). KATAN and KTANTAN: A family of small and efficient hardware-oriented block ciphers, in *Cryptographic Hardware and Embedded Systems—CHES 2009* (Lecture Notes in Computer Science Series), vol. 5747, C. Clavier and K. Gaj, Eds. Berlin: Springer-Verlag, 2009, pp. 272–288.

[27] S. Gordon, X. Huang, A. Miyaji, C. Su, K. Sumongkayothin, and K. Wipusitwarakun, "Recursive matrix oblivious RAM: An ORAM construction for constrained storage devices," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 12, pp. 3024–3038, 2017.

[28] L. Zhou, C. Su, W. Chiu, and K. H. Yeh, "You think, therefore you are: Transparent authentication system with brainwave-oriented bio-features for IoT networks," *IEEE Trans. Emerg. Topics Comput.*, 2017, to be published.

[29] A. Cavoukian. (2018). Privacy by design: The 7 foundational principles. [Online]. Available: https://www.iab.org/wp-content/IAB-uploads/2011/03/fred_carter.pdf

[30] N. Foukia, D. Billard, and E. Solana, "Pisces: A framework for privacy by design in IoT," in *Proc. 2016 14th Annu. Conf. Privacy, Security and Trust (PST)*, 2016, pp. 706–713.

[31] H. Wang, B. Sheng, and Q. Li, "Elliptic curve cryptography-based access control in sensor networks," *Int. J. Security Networks*, vol. 1, no. 3–4, pp. 127–137, Dec. 2006.

Liang Liu, Erik G. Larsson, Wei Yu, Petar Popovski,
Čedomir Stefanović, and Elisabeth de Carvalho

Sparse Signal Processing for Grant-Free Massive Connectivity

A future paradigm for random access protocols in the Internet of Things

INTERNET OF THINGS—ISTOCKPHOTO.COM/IAREMENKO
CIRCUITS—IMAGE LICENSED BY INGRAM PUBLISHING

The next wave of wireless technologies will proliferate in connecting sensors, machines, and robots for myriad new applications, thereby creating the fabric for the Internet of Things (IoT). A generic scenario for IoT connectivity involves a massive number of machine-type connections, but in a typical application, only a small (unknown) subset of devices are active at any given instant; therefore, one of the key challenges of providing massive IoT connectivity is to detect the active devices first and then decode their data with low latency. This article advocates the usage of grant-free, rather than grant-based random access schemes to overcome the challenge of massive IoT access. Several key signal processing techniques that promote the performance of the grant-free strategies are outlined, with a primary focus on advanced compressed sensing techniques and their applications for the efficient detection of active devices. We argue that massive multiple-input, multiple-output (MIMO) is especially well suited for massive IoT connectivity because the device detection error can be driven to zero asymptotically in the limit as the number of antennas at the base station (BS) goes to infinity by using the multiple-measurement vector (MMV) compressed sensing techniques. This article also provides a perspective on several related important techniques for massive access, such as embedding short messages onto the device-activity detection process and the coded random access.

Introduction

Wireless technology achievements in the past few decades are providing people with unprecedented connectivity, and there is growing interest in providing ubiquitous connectivity for machines and objects, many of which do not require interactions with humans [1]. This is being driven by the rapid advancement of the IoT, which will significantly impact the way we live our lives, the way we conduct business, deliver education, health care, and governmental services [2]. Typical IoT applications, as shown in Figure 1, include

1) smart health care in which the wearable devices transmit continuous streams of accurate data to the cloud for better care decisions

Digital Object Identifier 10.1109/MSP.2018.2844952
Date of publication: 28 August 2018

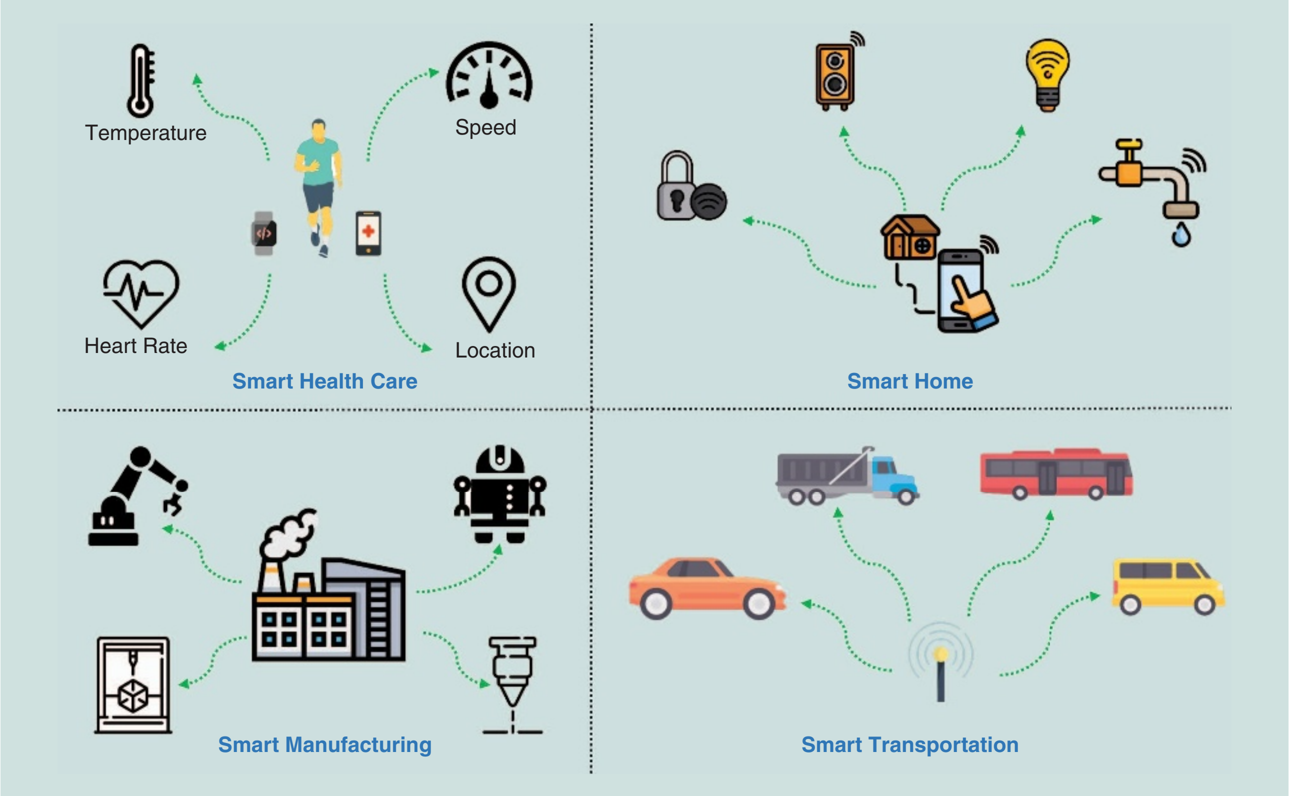


FIGURE 1. The applications of the IoT.

2) smart homes that enable home automation with the aid of intelligent appliances, such as the smart speaker even when the occupants are away from the home
3) smart manufacturing that supports streamlined business operations and optimized productivity in factories via automatically collecting and analyzing data from the sensors for making better-informed decisions to the actuators (e.g., robotics)
4) smart transportation in which the connected vehicles make transportation itself more efficient and help us get from place to place more quickly.

Targeting the emergence of the IoT, the fifth-generation cellular technologies road map has already identified massive machine-type communications (mMTC) as one of the three main use cases, along with enhanced mobile broadband and ultrareliable, low-latency communications.

The fundamental challenge of mMTC for the IoT is to enable data transmission from a massive number of devices in an efficient and timely manner. However, the key characteristic of IoT traffic is that the device-activity patterns are typically sporadic so that, at any given time, only a small and random fraction of all devices are active (Figure 2). The sporadic traffic pattern may be because devices are often designed to sleep most of the time to conserve energy and are only activated when triggered by external events, as is typically the case in a sensor network. In these scenarios, the active users must be dynamically identified along with the reception of their data, which is a challenging task.

Grant-based random access schemes

The common user access approach in cellular systems is to perform grant-based random access using the dedicated random-access control channel so that the uncoordinated devices can contend for physical-layer resource blocks for data transmission [3], as illustrated in Figure 3. In the first stage, each active device picks a random preamble, sometimes referred to as a *pilot sequence*, from a predefined set of orthogonal preamble sequences to notify the BS that the user has become active. In the second stage, the BS sends a response corresponding to each activated preamble as a grant for transmitting in the next step. In the third stage, each device that has received a response to its preamble transmission sends a connection request to demand resources for subsequent data transmission. In case a preamble has been selected by a single device, the connection request of the device is granted by the BS, which in turn sends a contention-resolution message informing the device of the resources reserved for the pending data transmission. However, if two or more devices have selected the same preamble in the first stage, their connection requests collide. When the BS detects a collision, it does not reply with a contention-resolution message; rather, the affected devices restart the random access procedure after a timer expires. In the above procedure, the messages sent by the active devices in the first and third phases correspond to metadata since they belong to control information for establishing the connection without containing any data information.

This access mechanism is an example of the classic ALOHA, which imposes a limit on the number of active devices that can

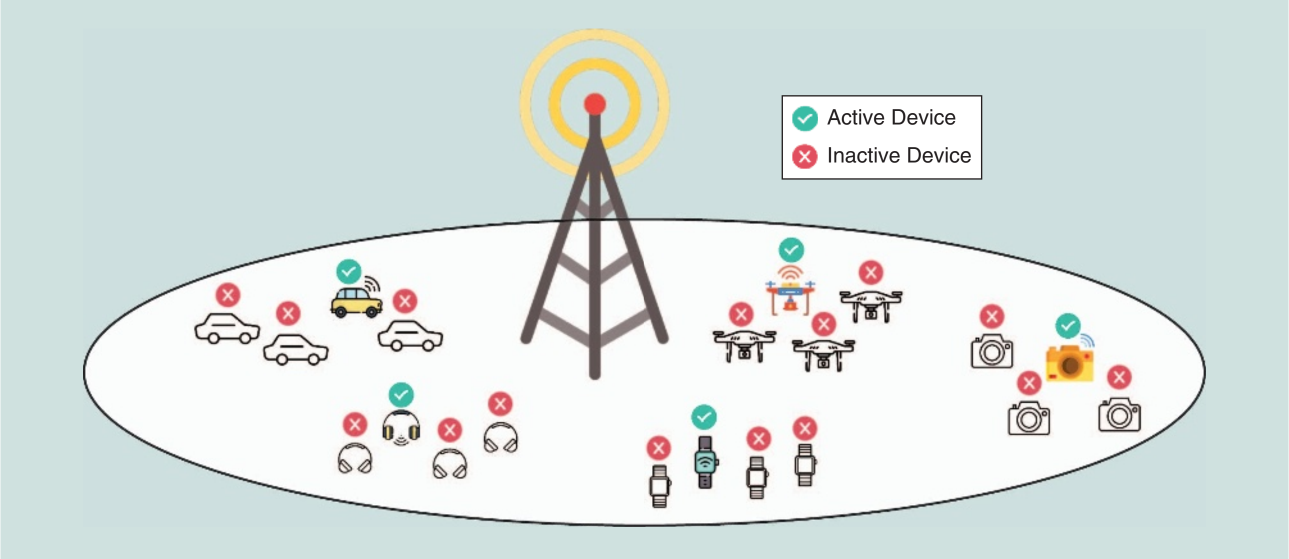


FIGURE 2. A typical IoT network with a massive number of devices, e.g., drones, smart watches, etc. Arising from the sporadic IoT data traffic, only a subset of devices in the network are active at each time slot.

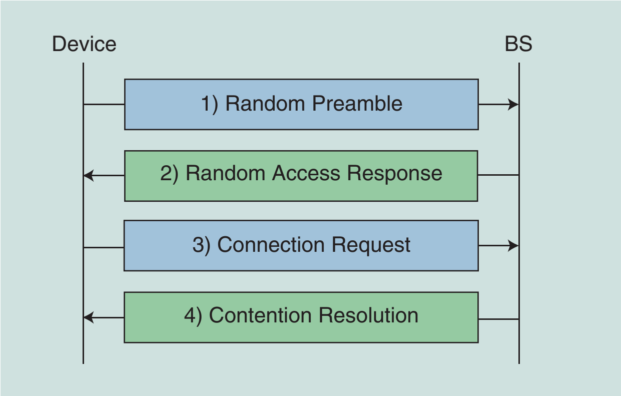


FIGURE 3. A grant-based random access procedure. Due to the lack of coordination, collisions occur when two or more devices select the same pilot, prompting the need for new access attempts in this case.

obtain the grant to access the network. Recently, extensive efforts have been devoted to different variations of the random access schemes with advanced contention-resolution strategies [4], [5]. However, due to the large number of collisions in the massive IoT scenarios, still many users cannot access the network even if some of the colliding connection requests could be resolved, as shown in Example 1.

Example 1

Consider a cellular network consisting of one BS and 2,000 users. Let L denote the length (and thus the number) of the orthogonal preambles available for the devices from which to choose. Assume that in each time slot 100 of these 2,000 devices are active, with each active device picking one of the L orthogonal pilots at random. The coherence bandwidth and the coherence time of the wireless channel are 1 MHz and 1 ms, respectively; therefore in each coherence block, 1,000 symbols can be transmitted. Moreover, we assume that both the scenario in which the contention resolution is not performed and the scenario in which (if) there is a collision, the BS can always grant access to one of the colliding devices (this could happen because of the capture effect in random access networks). Under this setup, the average numbers of devices that are granted permission to access the network for both the cases with and without contention resolution, versus the different values of L are plotted in Figure 4. The plot is obtained by Monte Carlo simulations. To guarantee a 90% success rate, at a minimum $L = 470$ and $L = 930$ out of 1,000 symbols are needed, respectively, as pilots for the cases with and without contention resolution.

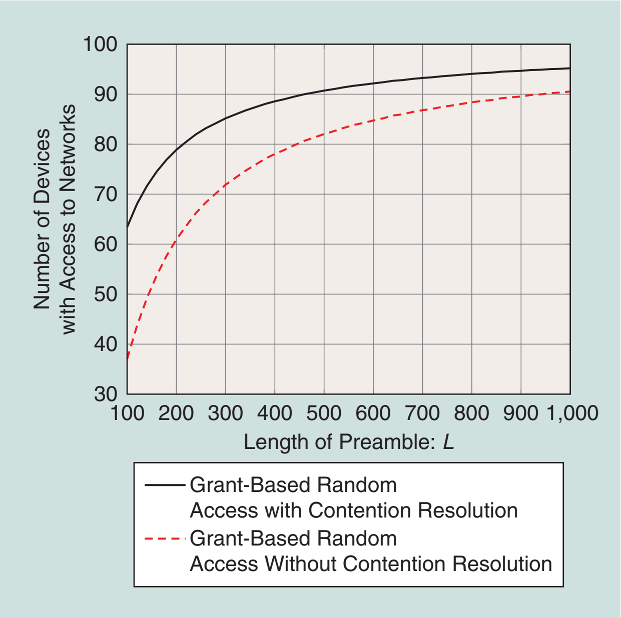


FIGURE 4. Grant-based random access with orthogonal pilots and captures. How many users can access the network?

A question arising from this example is how to accommodate more devices with low-latency requirements in the future massive IoT connectivity systems. One promising solution is the grant-free random access scheme based on the advanced compressed sensing techniques.

Grant-free random access schemes

Under the grant-free random access scheme, each active device directly transmits its metadata and data to the BS without waiting for any permission, as shown in Figure 5. In contrast to the grant-based random access scheme in which pilot sequences are randomly selected at each time slot, each device under the grant-free random access scheme is preassigned with a unique pilot sequence used for all of the time slots. This pilot sequence also serves as the ID for this user and is reminiscent of the role that the code-division multiple-access (CDMA) sequence plays in facilitating the extraction of a user data under interference from other users. At each time slot, the BS first detects the active devices by detecting which pilot sequences are used. Next, the BS estimates their channels based on the received metadata and then decodes the data with the estimated channels [6], [7].

The very fact that both metadata and data in the grant-free access are sent in a single step offers the possibility to decrease the access latency compared to the grant-based access. However, device-activity detection is now more challenging because it is not possible to assign orthogonal pilot sequences to all of the devices; this is due to the massive number of devices in the network as well as the limited channel coherence time. The difference with the classic CDMA systems is that the activation dynamics cover a much larger population, placing this problem in the realm of sparse signal processing.

This article aims to pave the way for a theoretical investigation on how the sparse signal processing technologies can enable accurate and efficient active device detection under the grant-free access scheme. We first point out that the device-activity detection can be cast into a compressed sensing problem. Next, a random pilot sequence design is introduced, and the use of an approximate message passing (AMP) algorithm [8] is proposed for detecting the active devices. We also demonstrate that massive MIMO [9], [10], which has already exhibited outstanding performance for enhancing the spectrum efficiency in human-type communications, provides an opportunity to leverage the so-called MMV compressed sensing technique [11], [12] to achieve asymptotically perfect device-activity detection accuracy in the massive IoT MTC. Another important fact about mMTC is that it relies primarily on short-packet transmissions. We elaborate on a new method to embed a small number of information bits in the short packets that can be decoded in the device-activity detection process. This is enabled by letting each active device randomly select one pilot from a predefined set and letting the BS detect which pilot is used by each active device using AMP. Finally, this article discusses the related technique of coded ALOHA [13] for device-activity detection.

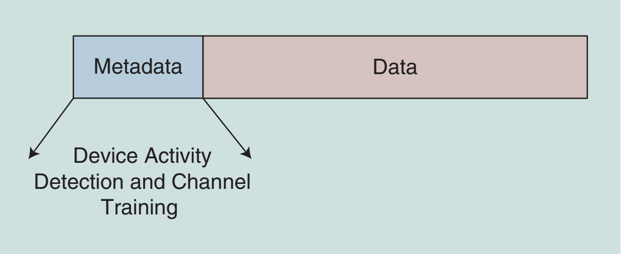


FIGURE 5. A grant-free transmission strategy. Metadata contains preamble for device-activity detection and channel estimation, and data is directly transmitted after metadata without waiting for the grant from BS.

Device-activity detection as a compressed sensing problem

As discussed in the previous section, it is the sporadic IoT traffic and device-activity detection that impose the greatest challenge to the design of the grant-free device access protocol. Interestingly, it is also the sporadic IoT traffic itself that provides a promising opportunity for tackling this challenge. As only a small subset of users is active at each time slot, user activity detection amounts to a sparse signal-recovery problem.

Suppose there are N users in the system, which are denoted by the set $\mathcal{N} = \{1,\ldots,N\}$. Furthermore, assume that the BS is equipped with one antenna, and the channel from user n to the BS is denoted by h_n. In each coherent time slot, define the user activity indicator function as

$$\alpha_n = \begin{cases} 1, & \text{if user } n \text{ is active,} \\ 0, & \text{otherwise,} \end{cases} \quad \forall n \in \mathcal{N}. \tag{1}$$

Assume that each device n decides in each coherence block whether to access the channel with probability ϵ_n in an independent manner. Then, α_n can be modeled as a Bernoulli random variable so that $\Pr(\alpha_n = 1) = \epsilon_n, \Pr(\alpha_n = 0) = 1 - \epsilon_n, \forall n$. As a result, on average, $K = \Sigma_{n=1}^{N} \epsilon_n$ devices are active in each time slot. The sparse activity level, ϵ_n, depends on the specific applications. The model is sufficiently general so that it can capture a variety of applications, e.g., a sensor fusion network in which the sampling rates at different sensors may even be different.

Suppose that each device n is assigned with one pilot sequence $\boldsymbol{a}_n \in \mathbb{C}^{L\times 1}$ with $\|\boldsymbol{a}_n\|^2 = 1$, where L denotes the length of device pilot sequence. We also assume that the active users are synchronized within the cyclic prefix so that the block-fading assumption yields a legitimate model for the channel. This is justified by having the BS send a beacon that invites uplink transmissions from the active devices. The received signal at the BS for device-activity detection is then

$$\boldsymbol{y} = \sqrt{\xi} \sum_{n \in \mathcal{N}} \alpha_n h_n \boldsymbol{a}_n + \boldsymbol{z} = \sqrt{\xi} \boldsymbol{A}\boldsymbol{x} + \boldsymbol{z}, \tag{2}$$

where $\boldsymbol{y} = [y_1,\ldots,y_L]^T \in \mathbb{C}^{L\times 1}$ is the received signals over L symbols, ξ is the total transmit energy of the pilot for each active device, $\boldsymbol{z} \in \mathbb{C}^{L\times 1} \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ is the independent additive white Gaussian noise (AWGN) at the BS, $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N]$ is the collection of pilot sequences of all the devices, and $\boldsymbol{x} = [x_1, \ldots, x_N]^T$, with $x_n = \alpha_n h_n$ denoting the effective channel of device n. The goal for the BS is to detect the active devices and detect their channels by recovering $\boldsymbol{x}$ based on the noisy observation $\boldsymbol{y}$.

Restricted by the limited coherence time in a practical massive IoT connectivity scenario, the length of device pilot sequence is much smaller than the number of devices, i.e., $L \ll N$. Hence, (2) describes an underdetermined linear system with more unknown variables than equations. However, since $\boldsymbol{x}$ is sparse with many zero entries based on (1), such a reconstruction problem is a sparse optimization problem that can be possibly solved via nonlinear compressed sensing techniques.

There are two main theoretical questions in compressed sensing: First, how can the sensing matrix $\boldsymbol{A}$ be designed to capture nearly all of the information about $\boldsymbol{x}$ with a minimal cost L? Second, given a sensing matrix $\boldsymbol{A}$, how can X be recovered from the noisy observation $\boldsymbol{y}$ even if $L < N$? In fact, these two questions are coupled: a good design of the sensing matrix $\boldsymbol{A}$ leads to an easier algorithm for recovering the sparse signal $\boldsymbol{x}$. For the massive IoT connectivity setting, this indicates that the device pilot sequences should be carefully designed to enable efficient activity detection schemes at the BS side.

Although a number of desirable properties for a good sensing matrix are known, e.g., restricted isometry property, optimizing the sensing matrix design remains a challenging problem. This article focuses on simple ways to construct the sensing matrix $\boldsymbol{A}$ that are easy to implement for practical pilot design. We consider how each entry of $\boldsymbol{A}$ is independent and identically distributed (i.i.d.) and randomly generated based on Gaussian distribution and review the AMP algorithm [8] to recover x [6], [14]. We also review other sensing matrix $\boldsymbol{A}$ choices and the corresponding compressed sensing algorithms, e.g., the sparse graph-based algorithm with a sparse $\boldsymbol{A}$ [15], and their applications in device activity detection, e.g., coded slotted ALOHA [13].

AMP-based device-activity detection

In the seminal work in [8], AMP is an efficient iterative-thresholding method designed for large-scale compressed sensing problems, which makes it appealing for massive IoT connectivity scenarios. An attractive feature of the AMP framework is that it allows an analytical performance characterization via state evolution [16]. Next, we describe how the AMP algorithm works for device-activity detection in massive IoT connectivity.

Device pilot sequence designs

In this section we assume that the entries of user pilots are generated from i.i.d. complex Gaussian distributions with zero mean and variance $1/L$, i.e.,

$$a_{n,l} \sim \mathcal{CN}(0, 1/L), \quad \forall n, l. \tag{3}$$

This particular choice of user pilot sequence is convenient for use with the AMP algorithm for two reasons: first, the convergence of the AMP algorithm for device-activity detection is guaranteed if $\boldsymbol{A}$ is generated in this way [8]; second, with a Gaussian-sensing matrix, the state evolution of the AMP algorithm is well established [16] based on which detection performance, e.g., missed-detection probability (probability that an active device is not detected) and false-alarm probability (probability that an inactive device is declared to be active) can be analytically characterized in the asymptotic limit.

Algorithm design and performance analysis

General form of an AMP algorithm

The AMP algorithm aims to provide an estimate $\hat{\boldsymbol{x}}(\boldsymbol{y})$ based on $\boldsymbol{y}$ that minimizes the mean-squared error (MSE)

$$\text{MSE} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\| \hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x} \right\|_2^2. \tag{4}$$

Based on an approximation of the message passing algorithm and starting with $\boldsymbol{x}^0 = \boldsymbol{0}$ and $\boldsymbol{r}^0 = \boldsymbol{y}$, the AMP algorithm proceeds at each iteration as [8] and [17]:

$$x_n^{t+1} = \eta_{t,n}((\boldsymbol{r}^t)^H \boldsymbol{a}_n + x_n^t), \tag{5}$$

$$\boldsymbol{r}^{t+1} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{t+1} + \frac{N}{L}\boldsymbol{r}^t \sum_{n=1}^{N} \frac{\eta'_{t,n}((\boldsymbol{r}^t)^H \boldsymbol{a}_n + x_n^t)}{N}, \tag{6}$$

where $t = 0, 1, \ldots$ is the index of the iteration, $\boldsymbol{x}^t = [x_1^t, \ldots, x_N^t]^T$ is the estimate of $\boldsymbol{x}$ at iteration t, $\boldsymbol{r}^t = [r_1^t, \ldots, r_L^t]^T \in \mathbb{C}^{L \times 1}$ denotes the corresponding residual, $\eta_{t,n}(\cdot): \mathbb{C} \longrightarrow \mathbb{C}$ is the so-called denoiser, and $\eta'_{t,n}(\cdot)$ is the first-order derivative of $\eta_{t,n}(\cdot)$. The basic intuition is that since the solution should minimize $\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \|^2$, the algorithm makes progress in (5) by moving in the direction of the gradient of $\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^t \|^2$, i.e., $(\boldsymbol{r}^t)^H \boldsymbol{a}_n$, $n = 1, \ldots, N$, and then promotes sparsity by applying an appropriately designed denoiser $\eta_{t,n}(\cdot)$. The residual is then updated in (6) and is corrected with an Onsager term involving $\eta'_{t,n}(\cdot)$.

State evolution

An important analytical result from the AMP algorithm is the so-called state evolution in the asymptotic regime when $L, K, N \rightarrow \infty$, while their ratios converge to some fixed positive values $N/L \rightarrow \omega$ and $K/N \rightarrow \epsilon$ with $\omega, \epsilon \in (0, \infty)$. In systems for massive IoT connectivity, these assumptions indicate that the length of the pilot sequence is in the same order as the number of active users or total users. After the tth iteration of the AMP algorithm, define a set of random variables $\hat{X}_n^t$'s as

$$\hat{X}_n^t = X_n + \tau_t V_n, \quad \forall n, \tag{7}$$

where the random variables X_n's capture of the distributions of x_n's, V_n follows the normal distribution, i.e., $V_n \in \mathcal{CN}(0, 1)$, and is independent of X_n as well as V_j, $\forall j \neq n$, and τ_t is the state variable, which changes from iteration to iteration as modeled by a simple scalar-iterative function

$$\tau_{t+1}^2 = \frac{\sigma^2}{\xi} + \omega\mathbb{E}\left[\left|\eta_{t,n}(X_n + \tau_t V_n) - X_n\right|^2\right]. \qquad (8)$$

Here, the expectation is over the random variables X_n's and V_n's over all of n. Under the aforementioned asymptotic regime, [16] shows that applying the denoiser to $(\boldsymbol{r}^t)^H \boldsymbol{a}_n + x_n^t$ in (5) is statistically equivalent to applying the denoiser to $\hat{X}_n^t$, as shown in (7).

The statistical model of AMP as given in (7) and (8) can be utilized to design the denoiser functions $\eta_{t,n}(\cdot)$'s in (5) and to quantify the performance of the AMP algorithm.

Minimax framework for denoiser designs

The flexibility in the AMP algorithm design lies in the denoiser $\eta_{t,n}(\cdot)$ in (5). In the AMP literature, the prior distribution of $\boldsymbol{x}$ is generally assumed to be unknown. In this case, the denoiser $\eta_{t,n}(\cdot)$ is designed under the minimax framework to optimize the AMP algorithm performance for the worst-case or least-favorable distribution of $\boldsymbol{x}$ [18]. Such a design leads to a soft-thresholding denoiser for promoting sparsity even for $\boldsymbol{x}$ with the worst-case distribution [8]:

$$\eta_{t,n}(\hat{x}_n^t) = \left(\hat{x}_n^t - \frac{\theta_n^t \hat{x}_n^t}{\left|\hat{x}_n^t\right|}\right)\mathbb{I}\left(\left|\hat{x}_n^t\right| > \theta_n^t\right), \qquad (9)$$

where the distribution of $\hat{x}_n^t$ is captured by $\hat{X}_n^t$, and $\theta_n^t > 0$ is the threshold for device n for the tth iteration of the AMP algorithm, which can be optimized based on the state evolution (8) to minimize the MSE as given in (4). With this denoiser, after the tth iteration of the AMP algorithm as shown in (5) and (6), device n is declared to be active if $\left|(\boldsymbol{r}^t)^H \boldsymbol{a}_n + x_n^t\right| > \theta_n^t$,and declared to be inactive otherwise. Note that AMP with soft-thresholding implicitly solves the least absolute shrinkage and selection operator (LASSO) problem [18], i.e., the sparse signal-recovery problem as an ℓ_1-penalized least squares optimization.

Bayesian framework for denoiser design

On the other hand, if the distribution of $\boldsymbol{x}$ is known in (2), we can design the minimum MSE (MMSE) denoiser via the Bayesian approach to minimize the MSE for the estimation of $\boldsymbol{x}$ as given in (4) [18]. Considering the equivalent signal model (7), the MMSE denoiser is given as the following conditional expectation:

$$\eta_{t,n}(\hat{x}_n^t) = \mathbb{E}\left[X_n \middle| \hat{X}_n^t = \hat{x}_n^t\right], \quad \forall t, n, \qquad (10)$$

where the expectation is over X_n and $\hat{X}_n^t$.

For example, if we assume a Rayleigh fading channel such that $h_n \sim \mathcal{CN}(0, \beta_n)$, where β_n denotes the path-loss and shadowing component of user n and is assumed to be known by the BS, then the effective channel $x_n = \alpha_n h_n$ follows a Bernoulli–Gaussian distribution. Under this particular distribution of $\boldsymbol{x}$, an analytical expression of the previously mentioned MMSE denoiser can be found in [14], which is generally nonlinear and has a complicated form. Similar to the soft-thresholding denoiser case, with the MMSE denoiser (10), we can detect the user activity based on whether the magnitude of $(\boldsymbol{r}^t)^H \boldsymbol{a}_n + x_n^t$ is larger than or smaller than a carefully designed threshold θ_n^t.

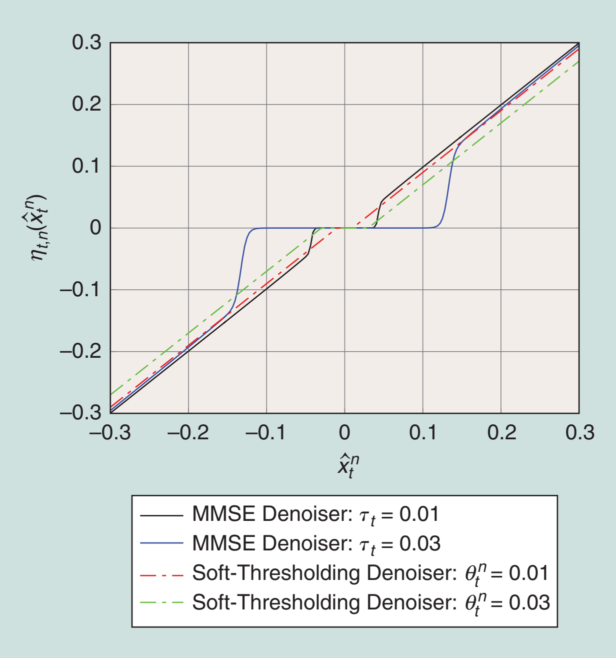


FIGURE 6. An MMSE denoiser versus a soft-thresholding denoiser in the AMP algorithm.

A comparison between the soft-thresholding and MMSE denoisers with a Bernoulli–Gaussian distributed $\boldsymbol{x}$ is given in Figure 6. The MMSE denoiser is also a thresholding-based denoiser, but softer around the regime near the threshold. Moreover, the threshold for the MMSE denoiser is obtained by calculating (10) to minimize the MSE (4), while the design of the threshold for the soft-thresholding denoiser follows a minimax framework, which is not optimal given the distribution of $\boldsymbol{x}$ in general.

Analytical performance characterization

The state evolution also allows an analytical performance characterization of the AMP algorithm. For example, with both the soft-thresholding and MMSE denoisers, a missed-detection event happens if one user is active but $\left|(\boldsymbol{r}^t)^H \boldsymbol{a}_n + x_n^t\right| < \theta_n^t$, while a false-alarm event happens if one user is inactive but $\left|(\boldsymbol{r}^t)^H \boldsymbol{a}_n + x_n^t\right| > \theta_n^t$. Since $\hat{x}_n^t$ defined in (7) captures the statistical distribution of $(\boldsymbol{r}^t)^H \boldsymbol{a}_n + x_n^t$, the probabilities of missed detections and false alarms for device n after the tth iteration of the AMP algorithm thus can be expressed as

$$P_{t,n}^{\text{MD}} = \Pr(\hat{x}_n^t < \theta_n^t \mid \alpha_n = 1), \qquad (11)$$
$$P_{t,n}^{\text{FA}} = \Pr(\hat{x}_n^t > \theta_n^t \mid \alpha_n = 0), \qquad (12)$$

respectively.

Given the distribution of $\boldsymbol{x}$ and denoiser $\eta_{t,n}(\cdot)$, we can track the values of τ_t's over all of the iterations based on the

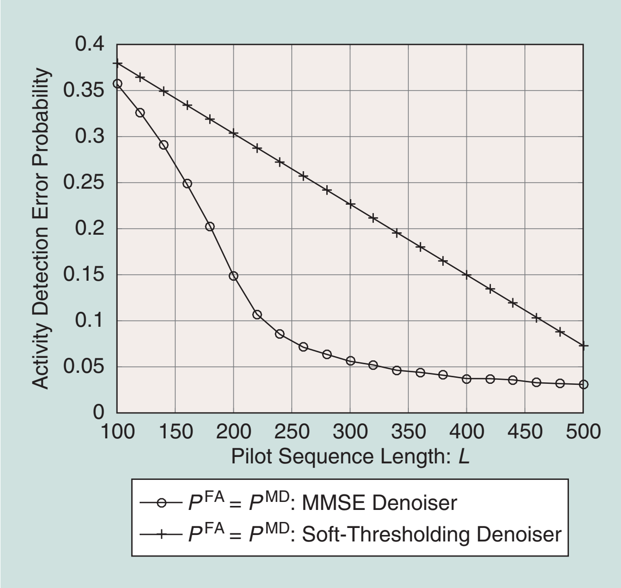


FIGURE 7. The probabilities of missed detections and false alarms versus pilot sequence length L.

state evolution (8), then calculate the probabilities of missed detections and false alarms based on (11) and (12).

Example 2

Here we provide a numerical example to show the probabilities of missed detections and false alarms achieved by the AMP algorithm under the same setup that is used in Example 1. The $N = 2{,}000$ devices are assumed to be randomly located in a cell with a radius of 1,000 m, while each device accesses the channel with an identical probability $\epsilon_n = 0.05, \forall n$, i.e., $\epsilon = 0.05$ and $K = 100$ of the $N = 2{,}000$ devices are active at any given time. The transmit power of each user for sending its pilot is $\rho^{\text{pilot}} = 23$ dBm. The power spectral density of the AWGN at the BS is assumed to be –169 dBm/Hz. Moreover, we define the system-level missed-detection and false-alarm probabilities as $P^{\text{MD}} = \Sigma_{n=1}^{N} P_n^{\text{MD}}/N$ and $P^{\text{FA}} = \Sigma_{n=1}^{N} P_n^{\text{FA}}/N$, where P_n^{MD} and P_n^{FA} denote the missed-detection and false-alarm probabilities of device n achieved by AMP after its convergence. Hence, $\epsilon N P^{\text{MD}}$ and $(1-\epsilon) N P^{\text{FA}}$ are the average numbers of missed-detection and false-alarm events at each time slot in a system with N devices. In addition, under both the soft-thresholding and MMSE-based AMP algorithms, the thresholds θ_n^t's are carefully selected so that $P^{\text{MD}} = P^{\text{FA}}$.

Figure 7 shows the device-activity detection accuracy achieved by the AMP algorithm with the soft-thresholding denoiser and the MMSE denoiser. With the MMSE denoiser, 90% active devices can be detected with the AMP algorithm when the length of pilot sequence satisfies $L > 220$. Recall that in Example 1 of the random access scheme, such a performance can be achieved only when the pilot sequence length is longer than 470 even if contention resolution is performed. Moreover, with a careful design of the MMSE denoiser, the MMSE denoiser-based AMP algorithm outperforms the soft-thresholding denoiser-based AMP algorithm in terms of device-activity detection.

From single-measurement vector to MMV: Massive MIMO for massive IoT connectivity

As compared to most other applications of compressed sensing such as imaging, a unique and essential opportunity provided by the wireless massive IoT connectivity system design lies in the potential for utilizing the MMV technique for compressed sensing [11], thanks to the multiantenna technologies used today in cellular networks. The previous section deals with the application of compressed sensing technique for user activity detection when the BS is equipped with one antenna. With regard to compressed sensing, the case with one measurement vector is referred to as a *single-measurement vector* (*SMV*) problem. As a revolutionary technology, massive MIMO has recently emerged for dealing with the future data deluge for human-type communications. In this section, we show that massive MIMO is also a natural solution for accommodating a huge number of IoT devices for future MTC. From the compressed sensing perspective, device-activity detection in massive MIMO systems corresponds to the MMV problem, which generalizes the sparse signal-recovery problem to the case with a group of measurement vectors for a group of signal vectors that are assumed to be jointly sparse and share a common support. It is of both theoretical and practical importance to investigate the role of massive MIMO on massive IoT connectivity.

Suppose that the BS is equipped with M antennas. In this case, the channel from user n to the BS is $\boldsymbol{h}_n \in \mathbb{C}^{M\times 1}$. Then, the signal model given in (2) is generalized to

$$\boldsymbol{Y} = \sqrt{\xi}\boldsymbol{A}\boldsymbol{X} + \boldsymbol{Z}, \tag{13}$$

where $\boldsymbol{Y} \in \mathbb{C}^{L\times M}$ is the matrix of received signals across M antennas over L symbols, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T$ with $\boldsymbol{x}_n = \alpha_n \boldsymbol{h}_n$ denoting the effective channel of user n, and $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M]$ with $\boldsymbol{z}_m \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \forall m$ is the independent AWGN at the BS.

As compared to the SMV signal model (2), the main difference lies in the fact that $\boldsymbol{X}$ in (13) is a row-sparse matrix, i.e., if one entry of one particular row of $\boldsymbol{A}$ is zero, the other entries of that row must be also zero. This information can be utilized to improve the user-detection accuracy. A comparison between the SMV model (2) and MMV model (13) is illustrated in Figure 8. Next, we discuss how to generalize the AMP-based algorithm shown in the "AMP-Based Device Activity Detection" section so that it can be used in the massive MIMO scenario. We also quantify the algorithm's significant improvement in device activity detection accuracy over the single-antenna BS case.

Algorithm design

With massive MIMO at the BS, the user's pilot sequence assignment still follows (3), which is the same as the case with one antenna at the BS. However, the AMP algorithm is modified as [12]

$$\boldsymbol{x}_n^{t+1} = \eta_{t,n}((\boldsymbol{R}^t)^H \boldsymbol{a}_n + \boldsymbol{x}_n^t), \tag{14}$$

$$\boldsymbol{R}^{t+1} = \boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}^{t+1} + \frac{N}{L}\boldsymbol{R}^t \sum_{n=1}^{N} \frac{\eta'_{t,n}((\boldsymbol{R}^t)^H \boldsymbol{a}_n + \boldsymbol{x}_n^t)}{N}. \tag{15}$$

As compared to (5) and (6), the dimensions of the signals are now $\boldsymbol{x}_n^t \in \mathbb{C}^{M\times 1}$ and $\boldsymbol{R}^t \in \mathbb{C}^{L\times M}$; moreover, the denoiser is a mapping in higher dimension, i.e., $\eta_{t,n}(\cdot) : \mathbb{C}^{M\times 1} \to \mathbb{C}^{M\times 1}$.

The state evolution of the AMP algorithm still holds for MMV in the asymptotic regime that $L, K, N \to \infty$ with fixed ratios $N/L \to \omega$ and $K/N \to \epsilon$. Specifically, define [12]

$$\hat{\boldsymbol{X}}_n^t = \boldsymbol{X}_n + \Sigma_t^{\frac{1}{2}} \boldsymbol{V}_n, \tag{16}$$

where the random vector $\boldsymbol{X}_n$ captures the distribution of $\boldsymbol{x}_n$, $\boldsymbol{V}_n \in \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$ is the independent Gaussian noise, and $\boldsymbol{\Sigma}_t$ can be tracked over iterations as

$$\Sigma_{t+1} = \frac{\sigma^2}{\xi}\boldsymbol{I} + \omega\mathbb{E}$$
$$\times\left[\left(\eta_{t,n}\left(\boldsymbol{X}_n + \Sigma_t^{\frac{1}{2}}\boldsymbol{V}_n\right) - \boldsymbol{X}_n\right)\left(\eta_{t,n}\left(\boldsymbol{X}_n + \Sigma_t^{\frac{1}{2}}\boldsymbol{V}_n\right) - \boldsymbol{X}_n\right)^H\right]. \tag{17}$$

Here, the expectation is over $\boldsymbol{X}_n$'s and $\boldsymbol{V}_n$'s over all of n. Then, in (14), applying the denoiser to $(\boldsymbol{R}^t)^H \boldsymbol{a}_n + \boldsymbol{x}_n^t$ is statistically equivalent to applying the denoiser to

$$\hat{\boldsymbol{x}}_n^t = \boldsymbol{x}_n + \Sigma_t^{\frac{1}{2}} \boldsymbol{v}_n, \tag{18}$$

where the distributions of $\hat{\boldsymbol{x}}_n^t$ and $\boldsymbol{v}_n$ are captured by $\hat{\boldsymbol{X}}_n^t$ and $\boldsymbol{V}_n$.

Based on the above state evolution (18), denoisers of the MMV-based AMP algorithm can be designed based on different criteria as for the SMV case. For example, the soft-thresholding denoiser is

$$\eta_{t,n}(\hat{\boldsymbol{x}}_n^t) = \left(\hat{\boldsymbol{x}}_n^t - \frac{\theta_n^t \hat{\boldsymbol{x}}_n^t}{\left\|\hat{\boldsymbol{x}}_n^t\right\|}\right)\mathbb{I}\left(\left\|\hat{\boldsymbol{x}}_n^t\right\|_2 > \theta_n^t\right), \tag{19}$$

Furthermore, assuming Bernoulli–Gaussian distributed $\boldsymbol{x}_n$'s, the MMSE denoiser

$$\eta_{t,n}(\hat{\boldsymbol{x}}_n^t) = \mathbb{E}\left[\boldsymbol{X}_n \middle| \hat{\boldsymbol{X}}_n^t = \hat{\boldsymbol{x}}_n^t\right] \tag{20}$$

is characterized in [6]. With both the soft-thresholding and MMSE denoisers, after the tth iteration of the AMP algorithm, user n can be declared to be active if $\left\|(\boldsymbol{R}^t)^H \boldsymbol{a}_n + \boldsymbol{x}_n^t\right\|_2 > \theta_n^t$, and declared to be inactive otherwise, where θ_n^t is the carefully designed threshold for device detection.

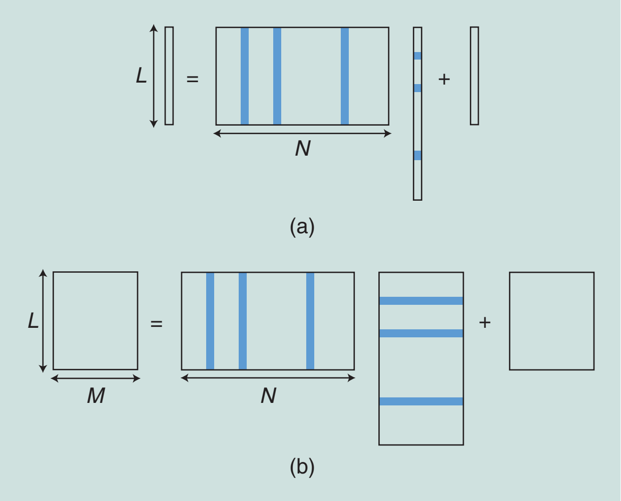


FIGURE 8. (a) An SMV-compressed sensing problem and (b) an MMV-compressed sensing problem. Group sparsity provides additional information about X because if one entry is zero, the other entries on the same row should be also zero.

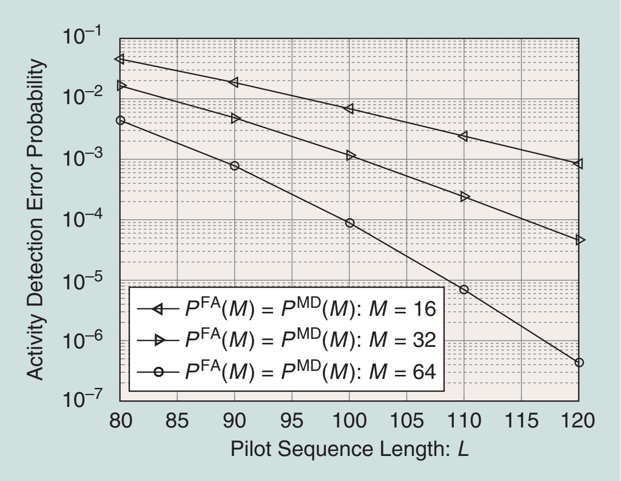


FIGURE 9. The probabilities of missed detections and false alarms versus pilot sequence length L with massive MIMO.

Asymptotically perfect device-activity detection

Fix the number of antennas at the BS, M, the missed-detection and false-alarm probabilities from the MMSE denoiser-based AMP algorithm, denoted by $P_{t,n}^{\mathrm{MD}}(M)$ and $P_{t,n}^{\mathrm{FA}}(M)$ (reducing to (11) and (12) when $M = 1$), are characterized in [6]. Interestingly, perfect device-activity detection is achieved in the asymptotic regime of $M \to \infty$ if the thresholds for device detection, i.e., θ_n^t's, are properly selected (c.f. [6, Th. 4]):

$$\lim_{M\to\infty} P_{t,n}^{\mathrm{MD}}(M) = \lim_{M\to\infty} P_{t,n}^{\mathrm{FA}}(M) = 0, \quad \forall t, n. \tag{21}$$

This important result implies that in a massive MIMO system, in which M can be larger than 100, the AMP-based grant-free access scheme is able to detect device activity with extremely high accuracy in the massive IoT connectivity systems.

Example 3

Here we provide a numerical example to show the power of massive MIMO for massive IoT connectivity under the same setup that is used in Examples 1 and 2. Figure 9 shows the probabilities of missed detections and false alarms (which are made equal by adjusting the detection threshold) versus pilot sequence length L, with $M = 16$, 32, and 64 antennas at the BS. Here, similar to Example 2, we define the system-level missed-detection and false-alarm probabilities as $P^{\mathrm{MD}}(M) = \Sigma_{n=1}^{N} P_n^{\mathrm{MD}}(M)/N$ and $P^{\mathrm{FA}}(M) = \Sigma_{n=1}^{N} P_n^{\mathrm{FA}}(M)/N$, where $P_n^{\mathrm{MD}}(M)$ and $P_n^{\mathrm{FA}}(M)$ denote the missed-detection and false-alarm probabilities of device n achieved by AMP after its convergence. As compared to Figure 7, it is observed that even with $M = 16$ antennas at the BS, both the missed-detection and false-alarm probabilities can be driven down to 10^{-3} when the pilot sequence is $L = 120$, several orders of magnitude lower than the SMV case with the same L.

This article mainly focuses on the device-activity detection performance under the grant-free access scheme. However, as shown in Figure 5, besides device-activity detection, channel estimation is performed as well via the metadata; moreover, data also should be decoded. Fortunately, the state evolution of the AMP algorithm enables us to characterize the channel estimation performance analytically, thus making it possible to quantify the user achievable rate with the effect of device-activity detection taken into consideration [7]. Readers interested in information-theoretical studies on the capacity of the massive IoT connectivity systems with randomly active devices (also known as a *many-access channel*) can refer to [19] and [20]. These references provide a justification for our proposed strategy to first detect the user activity via preambles then decode the user messages, i.e., the grant-free access scheme shown in Figure 5.

AMP-based device-activity detection with embedded information

In the "From Single-Measurement Vector to MMV: massive MIMO for Massive IoT Connectivity" section, the AMP algorithm was introduced for device-activity detection. In this section, we show how a modified version of AMP may be used for noncoherent detection of information bits embedded in the pilot transmission. As previously discussed, the two-phase grant-free access scheme shown in Figure 5 works very well for most of the cases when the user messages are of moderate and large size [7], [19], [20]. However, the strategy discussed in this section can be an effective alternative in the special case when very short messages (one or several bits) are transmitted.

Motivation

In many applications, the amount of data to be transmitted per block may comprise only a small number of information bits, or even a single bit. This situation is particularly common in control signaling, where the message may contain acknowledgment (ACK/NACK) bits in a retransmission protocol, or simply a concise request for a particular kind of response from the BS.

The transmission of extremely short packages is a challenging problem from two perspectives. First, fundamentally the protection of very short packets against transmission errors is very expensive. For a single bit, repetition coding is the only possible strategy and for short blocks, block codes with low-coding gains must be used. Second, as only error probability matters, capacity is an irrelevant metric. In fact, for extremely short blocks even finite-block-length information theory becomes inapplicable because the corresponding bounds and approximations are too loose to be of practical value.

As an aside, it is noteworthy that most academic work tends to deal with the transmission of long coded blocks, with Shannon capacity as the primary performance metric. Conversely, much of the effort invested in standardization and system design is concerned with the transmission of short data blocks on the control plane, for which Shannon capacity is mostly an illegitimate performance measurement. An explanation for this situation might be that digital transmission on the control plane is too hard to model and tackle with rigorous information theory: there is no Shannon theory available for its analysis; whereas, in contrast, established recipes are available for capacity analysis of long-block transmission. A contributing reason might also be that many academic researchers simply are unaware of the importance and the magnitude of the problem.

There are practical solutions for transmitting a single bit of control information. For example, [21] considers the joint transmission of linearly coded payload data and a single "additional bit." The transmitter uses the additional bit through a one-to-one mapping to select one of two possible codebooks for the encoding of the payload. The receiver uses a fast algorithm to detect the codebook that was used so that the additional bit can be detected before attempting to decode the payload data.

Algorithm design

In the context of grant-free random access with nonorthogonal pilots, the main focus of our discussion, a small number, say J, of bits $b_1, \ldots, b_J$ may be encoded as follows [22], [23]: Each terminal is assigned a priori 2^J distinct, typically nonorthogonal, pilots. Upon transmission, the terminal uses the bits $\{b_i\}$ to select one of these 2^J pilots; specifically, it selects pilot number $1 + b_1 + 2b_2 + 4b_3 + \cdots + 2^{J-1} b_J$, which, depending on the bits $\{b_i\}$, ranges from 1 to 2^J. The BS detects activity using the AMP algorithm; now, however, activity means the combination of the event that a particular terminal is active, and that a particular string of J bits is being communicated. One may think of the resulting communication scheme as noncoherent transmission.

The analytical model for device-activity detection with embedded information in a massive MIMO system is given by:

$$\boldsymbol{Y} = \bar{\boldsymbol{A}}\bar{\boldsymbol{X}} + \boldsymbol{Z}, \tag{22}$$

where $\bar{\boldsymbol{A}} = [\boldsymbol{a}_{1,1}, \ldots, \boldsymbol{a}_{1,2^J}, \ldots, \boldsymbol{a}_{N,1}, \ldots, \boldsymbol{a}_{N,2^J}] \in \mathbb{C}^{L \times 2^J N}$ denotes the collection of all the $2^J N$ pilots that can be used by the devices, and $\bar{\boldsymbol{X}} = [\bar{\boldsymbol{x}}_{1,1}, \ldots, \bar{\boldsymbol{x}}_{1,2^J}, \ldots, \bar{\boldsymbol{x}}_{N,1}, \ldots, \bar{\boldsymbol{x}}_{N,2^J}]^T \in \mathbb{C}^{2^J N \times M}$ denotes the collection of all the $2^J N$ effective channels of the devices. Specifically, the effective channel is modeled as $\bar{\boldsymbol{x}}_{n,i} = \alpha_{n,i} \boldsymbol{h}_n$, $n = 1, \ldots, N$ and $i = 1, \ldots, 2^J$, where

$$\alpha_{n,i} = \begin{cases} 1, & \text{if user } n \text{ is active and its } i\text{th pilot is used,} \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

As compared to (13) for sole device-activity detection, the dimensions of sensing matrix $\bar{\boldsymbol{A}}$ and effective channels $\bar{\boldsymbol{X}}$ are enlarged by a factor of 2^J to embed J bits information.

The AMP for device-activity detection in (14) and (15), in principle, could be directly applied to this problem as it stands. However, it is suboptimal because the BS knows a priori that among the 2^J pilots assigned to each terminal, only one can be active at a time, i.e., if $\alpha_{n,i} = 1$, then $\alpha_{n,j} = 0, \forall j \neq i$.

Here we discuss the modified AMP algorithm for joint detection of user activity and embedded information bits, as proposed in [22]. For conciseness of the exposition we focus on the case of a single embedded bit b (for which we omit the index), i.e., $J = 1$; then each user is assigned one of two unique, but generally nonorthogonal pilot sequences. The modification of the AMP should introduce the constraint that of the two possible pilots, at most, one may be transmitted at at a time; the possible options are that either none of these pilots are sent (device silent), the first one is sent (device active and communicates "0"), or the second one is sent (device active and communicates "1"). The overarching idea is to modify the AMP denoiser function $\eta_{t,n}(\cdot)$ to take this constraint into account.

In more detail, and similar to (16), let $\hat{\boldsymbol{x}}_{n,1} = \bar{\boldsymbol{x}}_{n,1} + \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{v}_n$ and $\hat{\boldsymbol{x}}_{n,2} = \bar{\boldsymbol{x}}_{n,2} + \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{v}_n$ be the two vectors associated with the two possible pilots (for information bit "0" and "1," respectively) for device n; we omit the iteration index t of the AMP algorithm here for brevity. The statistical characterization of $\hat{\boldsymbol{x}}_{n,1}$ and $\hat{\boldsymbol{x}}_{n,2}$ is

$$\hat{\boldsymbol{x}}_{n,i} \sim \begin{cases} \mathcal{CN}(\mathbf{0}, \beta_n \boldsymbol{I} + \Sigma), & \text{if } \alpha_{n,i} = 1, \\ \mathcal{CN}(\mathbf{0}, \Sigma), & \text{if } \alpha_{n,i} = 0, \end{cases} \quad i = 1,2. \tag{24}$$

Based on these characterizations, we construct the following likelihood ratios:

$$\begin{aligned} \lambda_{n,i} &= \frac{p(\hat{\boldsymbol{x}}_{n,i} \mid \alpha_{n,i} = 1)}{p(\hat{\boldsymbol{x}}_{n,i} \mid \alpha_{n,i} = 0)} \\ &= \frac{|\Sigma|}{|\beta_n \boldsymbol{I} + \Sigma|} \exp(-\hat{\boldsymbol{x}}_{n,i}^H((\beta_n \boldsymbol{I} + \Sigma)^{-1} - \Sigma^{-1})\hat{\boldsymbol{x}}_{n,i}). \end{aligned} \tag{25}$$

We now rework the denoiser so that in each time update the constraint is considered that, at most, one of the vectors $\bar{\boldsymbol{x}}_{n,1}$ and $\bar{\boldsymbol{x}}_{n,2}$ can be nonzero. Suppose that device n is detected to be active. In principle, a comparison of $\lambda_{n,1}$ and $\lambda_{n,2}$ to a threshold would yield a hypothesis test, that could be used to discriminate between the two possibilities $\alpha_{n,1} = 1$ and $\alpha_{n,2} = 1$ or equivalently $b = 0$ and $b = 1$; one of $\bar{\boldsymbol{x}}_{n,1}$ and $\bar{\boldsymbol{x}}_{n,2}$ could then be set to zero based on the outcome of this test. In this process, making a soft decision as given in (19), is instead preferable to avoid making premature incorrect decisions on the embedded bits, which may propagate to subsequent iterations. Experimentation in [22] demonstrates that a good heuristic is to use a soft decision obtained by taking the original soft-thresholding denoiser function given in (19) and multiplying the denoisers for $\bar{\boldsymbol{x}}_{n,1}$ and $\bar{\boldsymbol{x}}_{n,2}$ by $\gamma(\lambda_{n,1}/(\lambda_{n,1} + \lambda_{n,2}))$ and $\gamma(\lambda_{n,2}/(\lambda_{n,1} + \lambda_{n,2}))$, respectively, where $\gamma(x) = 1/(e^{-c(x-0.5)} + 1)$ is a modified sigmoid function with its inflection point at $x = 0.5$, where c is a parameter to control the sharpness of the sigmoid function. The theory is that the larger the likelihood ratio $\lambda_{n,1}$ is relative to $\lambda_{n,2}$, the more likely it is that $\alpha_{n,1} = 1$ or $b = 0$; the closer the weight for $\bar{\boldsymbol{x}}_{n,1}$ is to unity, the closer the weight for $\bar{\boldsymbol{x}}_{n,2}$ is to zero. Accordingly, the effect of the denoiser on $\bar{\boldsymbol{x}}_{n,1}$ is similar to the effect of the soft-thresholding (19) as used in the original AMP algorithm solely for device-activity detection, whereas on $\bar{\boldsymbol{x}}_{n,2}$ it is instead pushed down toward zero. A similar interpretation holds for the opposite case when $\lambda_{n,1} < \lambda_{n,2}$.

Importantly, while the modified denoiser outlined here yields good results in numerical experiments, it is not optimal in any known sense. Research opportunities are available to find improved denoisers that can make a better utilization of the constraint that at most one of $\bar{\boldsymbol{x}}_{n,1}$ and $\bar{\boldsymbol{x}}_{n,2}$ can be nonzero. An extension of the modified AMP denoiser to the case of multiple embedded bits is available [23].

A final remark is that the embedding of one or several information bit(s), of course, incurs the expense of storing more pilot sequences at the device and at the BS. Also, for a given coherence block length, more resources must be dedicated to pilot transmission to maintain the same error probability performance. Yet, the case of transmitting very short messages, the embedding scheme has been efficient compared to conventional schemes consisting of pilot-based channel estimation (using the sparsity/AMP-based techniques proposed in this article) followed by coherent detection [23].

Other compressed sensing techniques for device-activity detection

Aside from AMP, researchers with diverse backgrounds have developed many other powerful algorithms to reconstruct sparse signals from low-dimensional linear measurements, as given in (2) and (13). These compressed sensing algorithms can also be leveraged in our proposed massive IoT connectivity setting for device-activity detection, e.g., coded slotted ALOHA.

One powerful algorithm of low complexity is the sparse graph-based compressed sensing algorithm [15] in which the sensing matrix $\boldsymbol{A}$ is designed by sparsifying each row of the measurement matrix with zero patterns guided by sparse graph codes. The reason for such a sparse sensing matrix design is to disperse the signal into single tons that only contain one nonzero element in $\boldsymbol{x}$ and peel them off from multitons that contain two or more nonzero elements in $\boldsymbol{x}$ so that they can become single tons.

Figure 10 gives a simple example to briefly illustrate how this algorithm works to recover $\boldsymbol{x}$ in the ideal case without noise in (2), in which the dimensions of $\boldsymbol{x}$ and $\boldsymbol{y}$ are $N = 7$ and $L = 3$, respectively, and x_1, x_3, x_6 are nonzero entries in $\boldsymbol{x}$. Moreover, the sensing matrix is

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{26}$$

Due to the sparsity in both $\boldsymbol{A}$ and $\boldsymbol{x}$, in the received signal $\boldsymbol{y} = [y_1, y_2, y_3]^T$, y_1, y_2, and y_3 only contain information about x_1, $x_1 + x_3$ and $x_3 + x_5$, respectively. Thus, x_1 can be

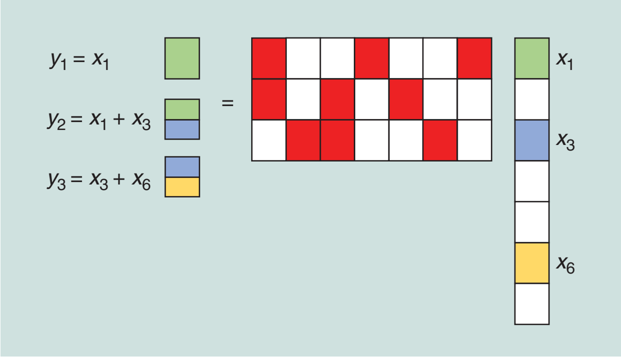


FIGURE 10. An example of the philosophy of the sparse graph-based compressed sensing algorithm [15]. The value x is a three-sparse signal in which the zero entries are white, while the nonzero entries, i.e., x_1, x_3, and x_6, are green, blue, and yellow, respectively. Moreover, the sensing matrix A is sparse, the zero entries are white, the nonzero entries are red, and x_1 is detected from y_1. Then, x_1 is subtracted from y_2 so that x_3 is detected. Similarly, x_6 is detected after x_3 is removed from y_3.

detected from the single-ton y_1. Next, x_1 is removed from y_2 and y_3, which become single tons so that x_3 and x_6 can be decoded. Note that the effect of the channels is not taken into consideration in this example.

Density evolution, a powerful tool in modern coding theory, tracks the average density of remaining edges that are not decoded after a fixed number of peeling iterations. The convergence of the graph-based compressed sensing algorithm is guaranteed by showing the convergence of the density evolution toward zero.

The above "successive interference cancellation" procedure is the principle of coded slotted ALOHA, a powerful multiuser access scheme in which the active devices transmit replicas of their packets in randomly chosen slots that contain both metadata (i.e., pilot sequences) and data. A successful detection of a packet replica in some slot enables removal of the related replicas from the slots in which they occur. This, in turn, lowers the number of colliding packets in the affected slots and boosts their detection probability, instigating new rounds of successive interference cancellation, and so on. If a single user packet can be detected in a slot, then the entries in $\boldsymbol{x}$ denote packets of active users, and entries in the sparse sensing matrix $\boldsymbol{A}$ denote the choice of the slots where the packets are repeated. On the other hand, the possibility to decode multiple user packets in a slot is also discussed in [13] and [24] to improve the detection performance.

Aside from AMP and the sparse-graph based algorithm, many powerful compressed sensing algorithms exist, including LASSO [25], orthogonal matching pursuit [26], and so on. Furthermore, the group sparsity in the MMV model (13) can also be utilized in LASSO [27] wherein the ℓ_1/ℓ_2 penalty, i.e., the sum of ℓ_2 norm penalty, is used to promote the desired sparsity pattern. The potential to apply these advanced compressed sensing techniques for user activity detection has been discussed in [28]–[30]. It would be of great significance to investigate which compressed sensing algorithm is best suited for device-activity detection in the massive IoT connectivity setting, in terms of the complexity of pilot sequence design, the pilot sequence length required to achieve reasonable device detection accuracy, the corresponding missed-detection and false-alarm probabilities performance, and channel estimation performance, and so on.

Conclusions

A key feature of the future IoT network is the massive number of devices, e.g., sensors, actuators, and so on, with sporadic data traffic. Facilitating the data transmission from so many IoT devices with extremely low latency poses plenty of new research challenges to the signal processing community. To embrace the upcoming era of IoT, this article advocates a grant-free access scheme that mitigates the delay arising from the contention resolution in the current random access scheme and outlines a compressed sensing-based approach for device-activity detection to enable the grant-free access scheme to work. Most notably, the massive MIMO technology, originally proposed for improving the spectrum efficiency of human-type communications, can boost the device-activity detection accuracy remarkably for massive IoT connectivity as well, with the aid of the MMV-based AMP algorithm. We have also discussed the potential of decoding some short messages along with the device-activity detection process.

Authors

Liang Liu (lianguot.liu@utoronto.ca) received his B.Eng. degree from Tianjin University, China, in 2010, and his Ph.D. degree from the National University of Singapore in 2014, where he is currently a research fellow in the Department of Electrical and Computer Engineering. He was previously a postdoctoral fellow at the University of Toronto, Canada. He was the recipient of the IEEE Signal Processing Society Young Author Best Paper Award in 2017, as well as the Best Paper Award from IEEE International Conference on Wireless Communications and Signal Processing 2011. His research interests include energy harvesting, convex optimization, and machine-type communications in fifth generation.

Erik G. Larsson (erik.g.larsson@liu.se) received his M.Sc. degree from Linköping University and Ph.D. degree from Uppsala University, both in Sweden. Currently, he is a professor at Linköping University. He coauthored *Fundamentals of Massive MIMO* in 2016 and *Space–Time Block Coding for Wireless Communications* in 2003. He is a member of the IEEE Signal Processing Society Awards Board and the *IEEE Signal Processing Magazine* editorial board. He received the IEEE Signal Processing Magazine Best Column Award in 2012 and 2014, the IEEE Communications Society Stephen O. Rice Prize in Communications Theory in 2015, the IEEE Communications Society Leonard G. Abraham Prize in 2017, and the IEEE Communications Society Best Tutorial Paper Award in 2018. He is a Fellow of the IEEE.

Wei Yu (weiyu@comm.utoronto.ca) received his B.A.Sc. degree in computer engineering and mathematics from the University of Waterloo, Canada, in 1997 and his M.S. and Ph.D. degrees in electrical engineering from Stanford University, California, in 1998 and 2002, respectively. He is a professor and the Canada research chair in information theory and wireless communications at the University of Toronto. He currently serves on the Board of Governors of the IEEE Information Theory Society and serves as chair of the IEEE Signal Processing Society's Signal Processing of Communications and Networking Technical Committee. He received the IEEE Signal Processing Society Best Paper Award in 2008 and 2017. He is a Fellow of the IEEE and a fellow of the Canadian Academy of Engineering.

Petar Popovski (petarp@es.aau.dk) received his Dipl.-Ing./Mag.-Ing. degrees in communication engineering from Ss. Cyril and Methodius University of Skopje, Republic of Macedonia, and his Ph.D. degree from Aalborg University, Denmark, where he is currently a professor. He received a European Research Council Consolidator Grant in 2015, the Danish Elite Researcher Award in 2016, the IEEE Fred W. Ellersick Prize in 2016, and the IEEE Stephen O. Rice Prize in 2018. He is currently an area editor of *IEEE Transactions on Wireless Communications* and a steering board member of the IEEE International Conference on Smart Grid Communications. His research interests include wireless communications/networks and communication theory.

Čedomir Stefanović (cs@es.aau.dk) received his Dipl.-Ing., Mr.-Ing., and Dr.-Ing. degrees in electrical engineering from the University of Novi Sad, Serbia, in 2001, 2006, and 2011, respectively. He is currently an associate professor in the Department of Electronic Systems, Aalborg University, Denmark. He is involved in several national and European Union projects related to the Internet of Things and fifth-generation communications. He is an editor of *IEEE Internet of Things Journal*. His research interests include communication theory and wireless and smart grid communications.

Elisabeth de Carvalho (edc@es.aau.dk) received her M.Sc. degree from Telecom Paris Sud in 1994 and Ph.D. degree in electrical engineering from Telecom Paris Tech in 1999. She was a postdoctoral fellow at Stanford University, California, and has worked in the field of digital subscriber line and wireless local area networks. Since 2005, she has been an associate professor at Aalborg University, Denmark, where she has led several research projects in wireless communications. She is a coauthor of *A Practical Guide to the MIMO Radio Channel*. Her research interests include signal processing for multiple-input, multiple-output (MIMO) communications with a recent focus on massive MIMO, including channel measurements, channel modeling, beamforming, and protocol aspects.

References

[1] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sept. 2016.

[2] L. Da Xu, W. He, and S. Li, "Internet of Things in industries: A survey," *IEEE Trans. Ind. Informat.*, vol. 10, no. 4, pp. 2233–2243, Nov. 2014.

[3] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, June 2013.

[4] O. Y. Bursalioglu, C. Wang, H. Papadopoulos, and G. Caire, "RRH based massive MIMO with on the 'fly' pilot contamination control," in *Proc. IEEE Int. Conf. Communications (ICC)*, 2016.

[5] E. Björnson, E. de Carvalho, J. H. Sørensen, E. G. Larsson, and P. Popovski, "A random-access protocol for pilot allocation in crowded massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2220–2234, Apr. 2017.

[6] L. Liu and W. Yu, "Massive connectivity with massive MIMO–Part I: Device-activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, June 2018.

[7] L. Liu and W. Yu, "Massive connectivity with massive MIMO–Part II: Achievable rate characterization," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2947–2959, June 2018.

[8] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18,914–18,918, Nov. 2009.

[9] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[10] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[11] J. Ziniel and P. Schniter, "Efficient high-dimensional inference in the multiple measurement vector problem," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 340–354, Jan. 2013.

[12] J. Kim, W. Chang, B. Jung, D. Baron, and J. C. Ye. (2011, Feb. 11). Belief propagation for joint sparse recovery. arXiv. [Online]. Available: http://arxiv.org/abs/1102.3289

[13] E. Paolini, C. Stefanović, G. Liva, and P. Popovski, "Coded random access: How coding theory helps to build random-access protocols," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 144–150, June 2015.

[14] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1890–1904, Apr. 2018.

[15] X. Li, S. Pawar, and K. Ramchandran, "Sub-linear time time compressed sensing using sparse-graph codes," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, 2015, pp. 1645–1649.

[16] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

[17] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, 2011, pp. 2168–2172.

[18] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. Motivation and construction," in *Proc. IEEE Information Theory Workshop*, 2010, pp. 1–5.

[19] X. Chen, T.-Y. Chen, and D. Guo, "Capacity of Gaussian many-access channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3516–3519, June 2017.

[20] W. Yu, "On the fundamental limits of massive connectivity," in *Proc. Information Theory and Applications (ITA) Workshop*, 2017.

[21] E. G. Larsson and R. Moosavi, "Piggybacking an additional lonely bit on linearly coded payload data," *IEEE Wireless Commun. Lett.*, vol. 1, no. 4, pp. 292–295, Aug. 2012.

[22] K. Senel and E. G. Larsson, "Device activity and embedded information bit detection using AMP in massive MIMO," in *Proc. IEEE Global Communications Conf. (Globecom)*, 2017.

[23] K. Senel and E. G. Larsson, "Joint user activity and noncoherent data detection in mMTC-enabled massive MIMO using machine learning algorithms," in *Proc. 22nd Int. ITG Workshop Smart Antennas (WSA)*, 2018.

[24] G. Wunder, C. Stefanović, P. Popovski, and L. Thiele, "Compressive coded random access for massive MTC traffic in 5G systems," in *Proc. 49th Asilomar Conf. Signals, Systems, and Computers*, Nov. 2015.

[25] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statistical Soc. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[26] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[27] L. Jacob, G. Obozinski, and J. Vert, "Group Lasso with overlap and graph Lasso," in *Proc. Int. Conf. Machine Learning*, 2009.

[28] Z. Utkovski, O. Simeone, T. Dimitrova, and P. Popovski, "Random access in C-RAN for user activity detection with limited-capacity fronthaul," *IEEE Signal Process. Lett.*, vol. 24, no. 1, pp. 17–21, Jan. 2017.

[29] H. Zhu and G. B. Giannakis, "Exploiting sparse user activity in multiuser detection," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 454–465, Feb. 2011.

[30] G. Wunder, H. Boche, T. Strohmer, and P. Jung, "Sparse signal processing concepts for efficient 5G system design," *IEEE Access*, vol. 3, pp. 195–208, 2015.

Wayes Tushar, Nipun Wijerathne, Wen-Tai Li, Chau Yuen,
H. Vincent Poor, Tapan Kumar Saha, and Kristin L. Wood

Internet of Things for Green Building Management

Disruptive innovations through low-cost sensor technology and artificial intelligence

INTERNET OF THINGS—ISTOCKPHOTO.COM/IAREMENKO
CIRCUITS—IMAGE LICENSED BY INGRAM PUBLISHING

Buildings consume 60% of global electricity. However, current building management systems (BMSs) are highly expensive and difficult to justify for small- to medium-sized buildings. The Internet of Things (IoT), which can collect and monitor a large amount of data on different aspects of a building and feed the data to the BMS's processor, provides a new opportunity to integrate intelligence into the BMS for monitoring and managing a building's energy consumption to reduce costs. Although an extensive literature is available on, separately, IoT-based BMSs and applications of signal processing techniques for some building energy-management tasks, a detailed study of their integration to address the overall BMS is limited. As such, this article will address the current gap by providing an overview of an IoT-based BMS that leverages signal processing and machine-learning techniques. We demonstrate how to extract high-level building occupancy information through simple, low-cost IoT sensors and study how human activities impact a building's energy use—information that can be exploited to design energy conservation measures that reduce the building's energy consumption.

Overview

Collectively, buildings are one of the major electricity consumers, representing 60% of total global electricity consumption. In the United States, for example, 70% of annual electricity use is due to buildings [1]. Such intense electricity usage by buildings is also true for many other countries (although detailed statistics may not be fully available due to lack of information or measurement). Therefore, there has been a significant push toward studying and developing ways to effectively manage electricity in buildings through efficient BMSs.

Current BMS solutions are, however, highly expensive and thus difficult to justify for use in small- and medium-size buildings. Additionally, due to a recent push for reducing electricity consumption and increasing operational efficiency, building managers need to deal with dynamic and diverse building requirements including anomaly detection, predictive maintenance, occupancy tracking, and electricity use optimization

Digital Object Identifier 10.1109/MSP.2018.2842096
Date of publication: 28 August 2018

with renewable integration. For example, in the United States, heating, ventilation, and air conditioning (HVAC) airflow is regulated by the U.S. Occupational Safety and Health Administration and is tied to maximum occupancy. Consequently, if no sensors are present, a room must be ventilated during normal working hours according to the maximum number of people who can be in the room (maximum seating capacity), thereby wasting considerable energy. From this perspective, sensing capabilities can lead to better situational awareness as well as more efficient, dynamic, and adaptive management of electricity and energy storage devices by incorporating intelligence into the BMS. The IoT has emerged as a promising solution to make this integration a reality.

Essentially, the IoT is a platform that connects devices over the Internet, allows them talk to one another and to humans, and, by doing so, enables the realization of desirable context-specific objectives such as energy savings (e.g., scheduling HVAC based on occupancy), condition monitoring (e.g., fault detection of HVAC), and predictive maintenance (e.g., the servicing of air filters in HVAC). Considering that the IoT is expected to change the future of smart BMSs, this article describes an IoT-based BMS that makes use of signal processing and machine-learning techniques. We note that signal processing has been employed extensively in wireless sensor networks for assisted living and information filtering. Therefore, it could be very helpful for extracting crucial information about building health using different sensors, as depicted in Figure 1. Based on this, we describe an energy-efficiency study conducted in a building test bed. The test bed is equipped with IoT devices and uses signal processing with machine learning to understand human activity and its impact on a building's energy use. To this end, the main contributions of this article are as follows:

- providing a literature review on the application of the IoT in building management as well as a discussion of the desired features of a smart BMS
- implementing machine-learning techniques in low-level devices, such as sensors and other IoT devices, via transfer learning and semisupervised learning techniques to enable IoT devices with low computation capability to perform machine-learning algorithms locally via edge computation
- illustrating how such techniques can benefit building management by providing useful information that can better characterize energy efficiency
- presenting some case studies from experiments in a real-world environment.

State of the Art

As buildings undergo years of use, their thermal characteristics deteriorate, indoor spaces get rearranged, and usage patterns change. In time, their inner and outer microclimates adjust to the changes in surrounding buildings, overshadowing patterns, city climates, and building retrofitting [2]. As a consequence, their performance frequently falls short of expectations. In this context, the IoT opens new opportunities to integrate intelligence into the BMS in a cost-effective manner through seamless integration of various sensors, smart meters, and actuators: the BMS can use these to monitor and identify different energy and environment-related parameters [3], analyze the health of a building, determine energy and thermal requirements, and, ultimately, determine the electricity usage behavior of different subsystems intelligently. The

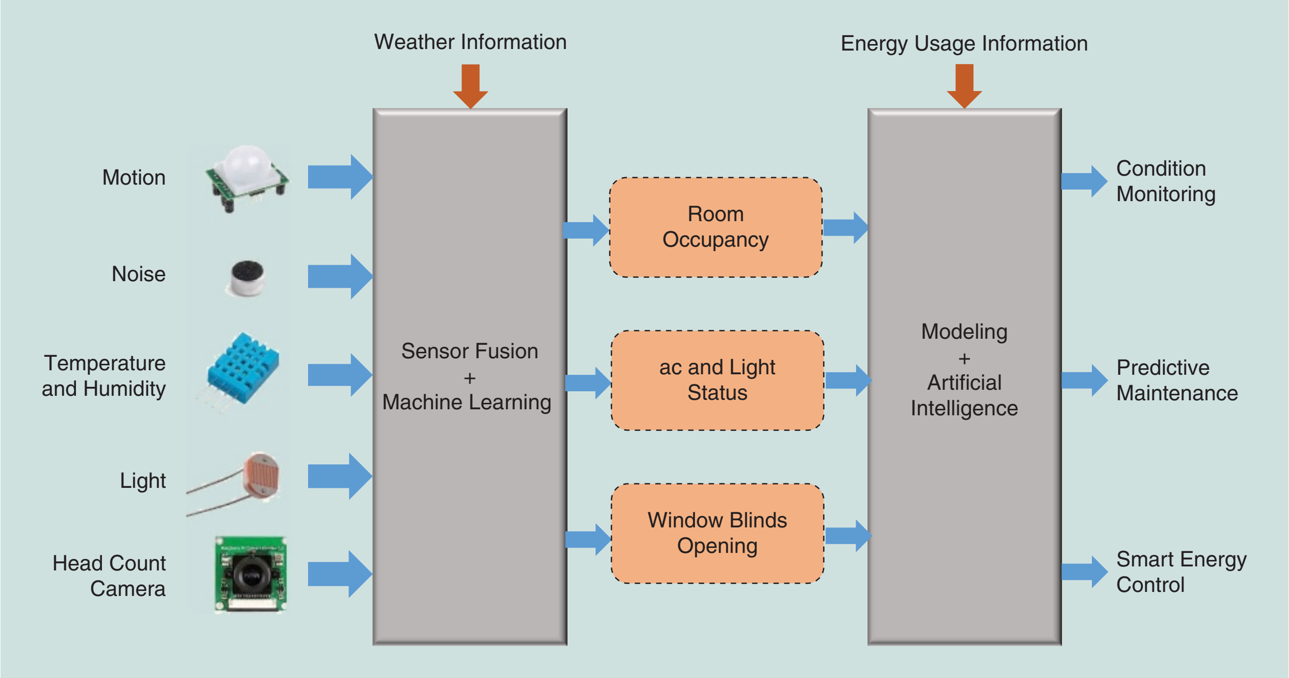


FIGURE 1. A demonstration of the use of sensor fusion and machine learning for IoT-based green building management.

performance of a BMS depends mainly on collecting large volumes of data from different building subsystems, which the BMS then analyzes and processes using various signal processing tools. Based on the application of IoT-based signal processing techniques in managing various subsystems within a building, existing studies can be divided into five general categories: 1) lighting, 2) HVAC, 3) flexible loads, 4) human detection, and 5) diagnostics and prognostics. We briefly describe these five categories in the following.

Lighting

Lighting accounts for a major fraction of global electricity consumption. In office buildings, for example, the electricity used for lighting can constitute up to 40% of total electricity consumption [4]. From this perspective, a number of studies have been conducted to develop solutions to help reduce electricity consumption based on a building's lighting. For instance, to achieve the desired illumination in a building having low electricity consumption, the authors in [4] propose a luminaire-based periodic sensor processing algorithm to implement a smart lighting control system. In [5], the authors present a Q-learning-based lighting control system that personalizes and employs users' perceptions of their surroundings as the feedback signal to better manage lighting intensity. In addition, [6] and [7] review lighting control techniques that use signal processing-based daylight-prediction and occupancy-detection methods, respectively.

HVAC

The HVAC system is another major consumer of electricity, accounting for 40% of total electricity consumption in U.S. buildings overall [1]. Consequently, developing some means to reduce the HVAC's electricity consumption has received considerable attention. In [8], the authors propose a Kalman filtering-based gray box model to predict and determine statistical process control limits for fault detection of HVAC systems. A similar signal processing technique is also used in [9] to control the power consumption of buildings without compromising the occupants' comfort level. An intelligent controller model is designed in [10] that integrates the IoT with cloud computing and web services; in addition, the authors develop wireless sensor nodes to monitor the indoor environment and HVAC inlet air as well as a wireless base station to control the HVAC actuators. Lastly, [11] proposes a smart home energy-management system using the IoT and big data analytics, predominantly focused on the HVAC system's electricity consumption; in particular, the proposed mechanism makes use of off-the-shelf business intelligence and big data analytics software packages to better manage energy consumption and meet consumer demand. Other examples of such studies in the context of HVAC can be found in [12] and [13].

Flexible loads

The third area of study focuses on monitoring and controlling electricity consumption by other flexible loads within buildings. Examples of such loads include washing machines, dishwashers, ovens, electric vehicles, and energy storage systems. IoT devices can effectively monitor the operational status of these loads and exploit signal processing techniques to predict their usage patterns and effectively control their operation for better energy management (or demand response). For example, a learning-based signal processing tool for demand management is designed in [14], and a deep-learning-based signal processing approach is implemented in [15] for nonintrusive monitoring of all loads in an entire building. In [16], the authors design a long- and short-term memory for load forecasting based on residential-behavior learning using recurrent neural networks, while, in [17], accurate indoor occupancy tracking within a building is implemented using multisensor fusion.

Human detection

In the area of human detection for BMSs, a number of machine-learning techniques have been used for head count and occupancy-detection purposes. For example, parametric and nonparametric algorithms (including background subtraction models and Gaussian processes [18]) have been used with a camera for head count; these algorithms are implemented using the OpenCV library. Further, for occupancy detection, thermal imaging [19], pyroelectric infrared sensors [20], and red-green-blue camera-based techniques have been used extensively. In addition, the authors of [21] present an example of sound-sensor-based applications for occupancy detection that use high sampling rates to classify activities.

Diagnostics and prognostics

The HVAC system is the most complex system and greatest energy consumer in most buildings. Faulty equipment within the HVAC system leads to inefficient system operation. Hence, keeping such systems in good operational condition is important. However, regular maintenance of HVAC systems is time consuming. Given this context, IoT devices allow us to develop advanced predictive-maintenance, fault-detection, and diagnostics applications for these systems. For instance, using data collected from IoT devices, a data-driven fault-detection method is developed in [8].

Based on the discussion here (summarized in Table 1), it is clear that the IoT makes possible numerous useful applications in designing smart and efficient buildings. Nevertheless, obtaining desirable BMS performance by applying different signal processing techniques depends significantly on the actual IoT devices that monitor and collect large amounts of data in respective contexts and then feed these data to the processor. Although, as noted previously, studies are available separately on IoT-based BMSs and applications of signal processing techniques for some aspects of building energy management, studies focused on their integration for overall building energy operation in practical settings are limited. We address this lack of integration by providing an overview of how IoT devices can be coupled with signal processing techniques to better understand a building's electricity usage performance.

Table 1. A summary of the surveyed literature on applying IoT-based signal processing for BMSs.

Category	Main Focus of the Study	Adopted Technical Approach	Surveyed Studies
Lighting	Use sensor-based data to shape the output of light-emitting-diode lighting systems to achieve desired illumination conditions and lower electricity consumption within a building	Proportional-integral-derivative control, custom-built android mobile applications, controller optimization	[4]–[7]
HVAC	Reduce electricity consumption by the HVAC system of a building without affecting the privacy and comfort level of the building's occupants	Kalman filtering, gray box models, cloud computing, big data analytics, business intelligence models	[1], [8]–[13]
Flexible loads	Monitor and schedule electricity consumption by flexible loads within a building to reduce electricity costs	Machine learning, deep learning, recurrent neural networks, behavioral modeling	[14]–[17]
Human detection	Monitor the number of people (or occupancy detection) in a room within a building to facilitate energy-consumption modeling	Background subtraction models, Gaussian processes, sound-sensor-based approaches	[18]–[21]
Diagnostics and prognostics	Advance predictive maintenance, fault detection, and diagnostics applications of systems using IoT-based data	Kalman filtering, gray box models	[8]

Trending technologies for BMS

Here, we provide an overview of trending technologies being used for the effective design and development of a BMS data acquisition, management, and control platform. These technologies treat the BMS as a cloud-based ecosystem that 1) uses social interaction among the people within communities to establish sophisticated global behavior patterns that can achieve different social, financial, and scientific goals; 2) uses green energy and storage resources for environmental sustainability; and 3) affirms the overall establishment of a smart system with autonomous decision-making capability.

IoT for seamless integration and processing

At present, it is difficult and expensive for a building manager to be fully aware of a building's health in real time, due to current buildings' limited sensing and control capabilities. Hence, the design of an integrated data-acquisition and control system based on an open architecture and a cloud-enabled IoT can help reduce the cost of setting up a BMS. An integrated IoT system allows the building manager to monitor and sense the building's different environmental parameters (e.g., through motion and noise detectors, temperature and humidity sensors, and electricity and water flow meters), collect the relevant human activity information (occupancy, heat map, etc.), and estimate the energy usage (e.g., by comparing the current information with previously collected historical data), which will be fed into a smart management system that will manipulate actuators (e.g., switches, controllers, and thermostats) to efficiently manage the building's environment according to expectations and designated rules.

Such an IoT platform (which is an open platform) can interface and connect with various subsystems of different vendors, e.g., sensing subsystems (people counting, temperature, humidity, light, noise, and motion), control subsystems (thermostats, switches, smart plugs, and actuators), and metering subsystems (energy consumption, water flow, etc.). Currently, various off-the-shelf products and systems are available for the IoT; for example, utility use worldwide is trending toward smart energy profiles, such as batteryless energy harvesting switches (e.g., Enocean), low-cost Wi-Fi controllers (e.g., Particle) and thermostats (e.g., Nest by Google), and many others. As these technologies advance, we expect that more and more IoT devices will be available on the market. Hence, there are great opportunities to tap the capability of these growing IoT systems. Nevertheless, to implement an IoT system, one needs to address the challenges of scalability, flexible provisioning, interoperability, and low latency [22].

Attribution of energy usage to human activities

One of the key technologies receiving considerable attention for deployment as a part of a BMS is the integration of human activity tracking within the control system as a way to understand how a building's electricity usage is affected by the number of occupants and their various activities. By "tracking human activity," we refer to the tracking of human movement within a designated area, which can be accomplished using a smartphone scanning sensor. Essentially, such a sensor can scan smart devices in the vicinity (smartphones, mobile tablets, and the like) and, at the same time, record the smart device's duration of stay and media access control address. This determines not only the heat map (i.e., human count) of the designated area but also presents information concerning the duration of stay and the movement path—and, potentially, social relationships as well. Such tracking can provide building managers with information on occupants' efficient use of different areas in a building and help them in recommending further modification (architectural or electrical system) of a space if necessary for greater overall energy efficiency.

Big data analytics for insightful analysis

Big data management is, in essence, the core software layer that ultimately drives the BMS through big data analytics, including prediction (e.g., predictive maintenance), model building, complexity mapping, and visualization. Big data analytics aids the dynamic management of energy consumption via monitoring and analyzing energy-related activities to minimize unintended energy and water consumption. This is basically a process of examining the large data set obtained through the IoT. Using this process, a building manager can identify information on human activity, weather conditions, the microclimate within the building, and related energy wastage. Doing so requires designing algorithms that can accurately extract the intercorrelations among load consumption, human occupancy and movement activity, energy wastage, renewable energy generation, and weather conditions, allowing the creation of effective models from real-time large-scale data sets to perform predictive maintenance.

Renewable energy and storage for increasing the flow of green energy

To facilitate the flow and use of green energy, BMSs also explore possible provisioning of renewable energy resources in buildings through dynamic scheduling and control. In particular, the BMS first collects data on weather conditions and subsequent renewable energy generation through the low-cost IoT system. Then, together with the information about human activity within the building, the BMS exploits big data analytics to determine how to optimally schedule the dispatch of renewable energy from its distributed sources as well as the charging and discharging of the respective storage devices.

For example, a solar thermal system integrated with hot water storage is becoming very popular in commercial buildings. Based on the previously discussed process (i.e., using the available solar generation for heating water to meet demand based on human activity), a building's BMS can use its big data analytics to predict how much hot water will be needed for the building. Thus, it can dynamically schedule the heat pump to turn on to heat hot water based on the availability of renewable energy.

IoT-based human activity detection and building efficiency

As mentioned earlier, understanding human activity is particularly important to energy efficiency. In this section, we provide a brief case study to illustrate how low-cost IoT devices can be used along with signal processing and machine-learning techniques to understand the number of people in a particular area of a building and provide building management with key insights for effectively managing the building's electricity consumption.

Head counting with an overhead camera

To realize a low-cost head count camera, we use a camera with a fish-eye lens that captures images at 30 frames/s. The camera is installed directly on top of the entrance door of the selected room. The images captured by the camera are processed locally. Only the head count number is uploaded to the cloud. For head counting, we explore a number of signal processing techniques including the following.

OpenCV image processing

We first use the OpenCV library with traditional image-processing techniques for head counting. To overcome the influence of moving objects such as a door, the background subtraction method is fused with the color detection method in head count detection. Unfortunately, the accuracy of this method becomes unacceptable when the light inside the room is switched off. This motivates us to use deep-learning techniques in the head count camera.

Motivation for transfer learning

Recently, deep learning—and, especially, convolutional neural networks (CNNs)—has made great progress in object recognition. However, building an object-recognition model with CNNs is tedious due to the significant amount of data and the resource requirements for training purposes. For instance, the model for the ImageNet Large Scale Visual Recognition Challenge was trained on 1.2 million images over a period of 23 weeks in multiple graphics processing units. As a consequence, it has become popular among researchers and practitioners to use transfer learning and fine-tuning (i.e., transferring the network weights, which have been trained on a rich data set, to the designated task, e.g., detecting people on images in this study). In particular, we use the pretrained single shot detection (SSD) multibox model [23] as the network due to its higher accuracies, high frame rate, and suitability for embedded application.

SSD multibox

SSD matches objects with default boxes having multiple feature maps with different aspect ratios. Each element of the feature map has either four or six default boxes associated with it. Any default box having a Jaccard overlap higher than a threshold of 0.5 with a ground truth box is considered a match. SSD has six feature maps in total, each responsible for a different scale of objects, thus allowing it to identify objects across a large range of scales. SSD runs 3 × 3 convolution filters on the feature map to classify and predict the offset to the default boxes.

Transfer learning and fine-tuning

There are two main procedures we adopt when using the pretrained SSD multibox model. The first is transfer learning. We use an SSD model that is pretrained for the Microsoft Common Objects in Context (COCO) data set [24]. Then, we change the number of classes in the box predicator according to our requirement. Second is fine-tuning. We keep the pretrained weights of the feature extractor and use them as initial values for retraining. Note that, because the feature extractor is trained for a large rich data set over millions of iterations, the weights are stable and converged. As such, we only fine-tune the feature extractor weights according to our data set.

In Figure 2, we depict results based on OpenCV libraries using only our pretrained SSD model and transfer learning. Figure 2 also demonstrates a frame captured under low lighting conditions. One can observe that OpenCV and the pretrained SSD model do not show good results under low lighting conditions. As illustrated in Figure 2(b), the OpenCV method captures the same person as two instances, and the pretrained SSD model does not capture anything at all [Figure 2(c)]. Nonetheless, the transfer learning method is well generalized for both good and low lighting conditions [Figure 2(d)] and can be used for the people-counting algorithm.

It is important to note that, to reduce the processing complexity and power consumption for occupancy detection, the

MobileNet architecture is used as the feature extractor for the SSD model. The input image size was restricted to 250 × 250 pixels without losing accuracy, which greatly reduces the computational power. Further, while the original pretrained SSD MobileNet can detect 99 object classes accurately, we reduce the number of classes to two. This helps lower the power consumption of the module without limiting the performance of the detection (because our intention is to detect only person versus nonperson cases).

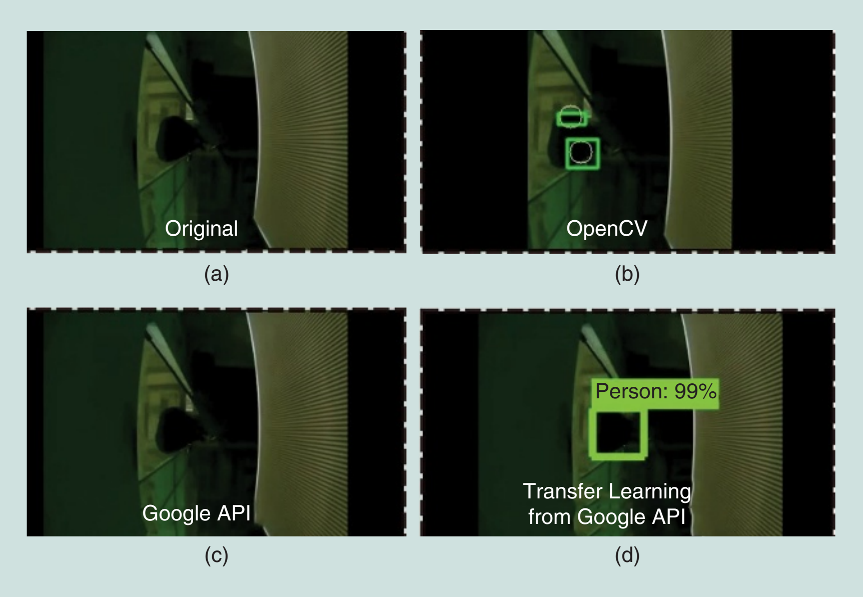


FIGURE 2. Detection results using four different methods in low lighting conditions: (a) the original frame, (b) using OpenCV libraries, (c) using the SSD model trained for the COCO data set, and (d) using transfer learning. API: application programming interface.

Occupancy detection with a sound sensor

Because a motion sensor has limited range, we also explore the suitability of using a sound sensor for occupancy detection. We are motivated to explore sound sensors because the visual approach using a camera can capture occupants' identity and record their activities within a selected space of the building—which not only violates user privacy but also requires a very large storage space and data rate for processing in real time. As such, existing studies of occupancy detection in building environments, such as [25] and [26], have used techniques that exploit environmental information (including carbon dioxide level, noise level, humidity level, and particulate matter concentration) instead of using a camera. From this perspective, an acoustic approach using a sound sensor is a potential alternative being explored here.

A sound sensor is typically used to detect the loudness in ambient conditions, with an input taken from a microphone and amplified. We use a low-cost analog sound sensor with a signal-to-noise ratio of 55 dB at 1-kHz maximum input frequency. Note that, due to different room structures (room size, wall material, furniture, etc.), sensors are placed at different locations in different rooms, which results in a collection of different noise levels. The sensor data are sampled every 100 ms. Based on our empirical research and previous study [27], a sampling rate of 10 Hz is enough to distinguish human activities within the environment. In addition, such low-rate sampling can help reduce the sensors' energy consumption and computation without losing necessary information.

To accurately detect human occupancy in selected areas, we explore four different techniques: 1) the threshold method, 2) the unsupervised learning through clustering method, 3) the supervised learning through deep-learning method, and 4) a semisupervised learning technique that employs the deep classifier to label the unsupervised results.

Thresholding method

We accumulate the sound-sensor data for every 5-min period and measure an empirical threshold value for each room to determine occupancy. We observe that threshold values are different for different rooms and sometimes even for different days. As Figure 3 shows, the appropriate threshold for human activity detected for room P04 is 8,000, while the threshold becomes 12,000 for room P02. Here, P02 and P04 are two selected rooms of the test bed (Figure 4). Therefore, simply using the threshold method is not robust. Further, it is tedious to calibrate and find the optimal threshold for each individual room.

Unsupervised learning using clustering

To achieve robust occupancy detection using a noise sensor with no calibration, we employ unsupervised learning. Instead of sending just the accumulated noise value, we send the noise histogram for each 5-min interval according to the following arrangement:

- bin 1 (sample values range: 0–6)
- bin 2 (sample values range: 6–10)
- bin 3 (sample values range: 10–15)
- bin 4 (sample values range: 15–30)
- bin 5 (sample values range: 30–50)
- bin 6 (sample values range: 50–75)
- bin 7 (sample values range: 75–100)
- bin 8 (sample values range:100 and above).

The noise histogram refers to a set of bins (ranges), in which each bin represents a specific range of data collected by the sound sensors. While these eight levels of the histogram are used in our case, we believe four or even fewer levels could also be sufficient.

We adopt a similar unsupervised learning process in our previous work [27] in an outdoor smart city environment. To generate meaningful features of the histogram data, we use the localized behavior of wavelet transformation; the Haar basis function is used as the mother wavelet, due to its own discontinuous nature (the Haar basis function is a mathematical function that generates the feature set for the histogram's representation of data). We represent each 5-min histogram in terms of the Haar basis function. Further, principal component analysis is used to remove the high correlation between histogram bins. In addition, we remove the redundant noise in the data by using only the *n* principal components that capture 95% of the variation in the data. Moreover, we use hierarchical clustering and calculate the optimal number of clusters using the Calinski-Harabasz index.

Supervised learning through deep learning

A major issue in terms of the unsupervised learning technique is that it is up to a human to interpret the outcome. For example, in our case, we may get three to five different clusters as the output of the unsupervised learning, as shown in Figure 3 (four clusters for room P02 and five clusters for room P04). In some cases, one cluster represents the unoccupied duration; in other cases, two clusters may represent the unoccupied duration. Because the unsupervised learning method does not provide any meaningful understanding of the clusters, we employ a deep neural network (DNN) classifier, using the thresholding method with a "reasonable threshold" as the "ground truth." The output of the classifier is the probability of occupancy. For instance, if the probability of occupancy is greater than 0.5, we determine that the space is occupied.

To this end, we first use a sparse autoencoder to extract meaningful features for histograms. When training for feature extraction, we use only histograms related to one class of data (i.e., only those histograms related to the occupied class) to train the autoencoder rather than using the data of both classes (i.e., occupied and unoccupied). By doing so, we expect to construct more distinguishable features for histograms.

When training the classifier, we fix the weights of the pretrained autoencoder and thus optimize only the weights of the DNN. To do so, we first use the sparse autoencoder to construct sparse features for the histogram and then use these features

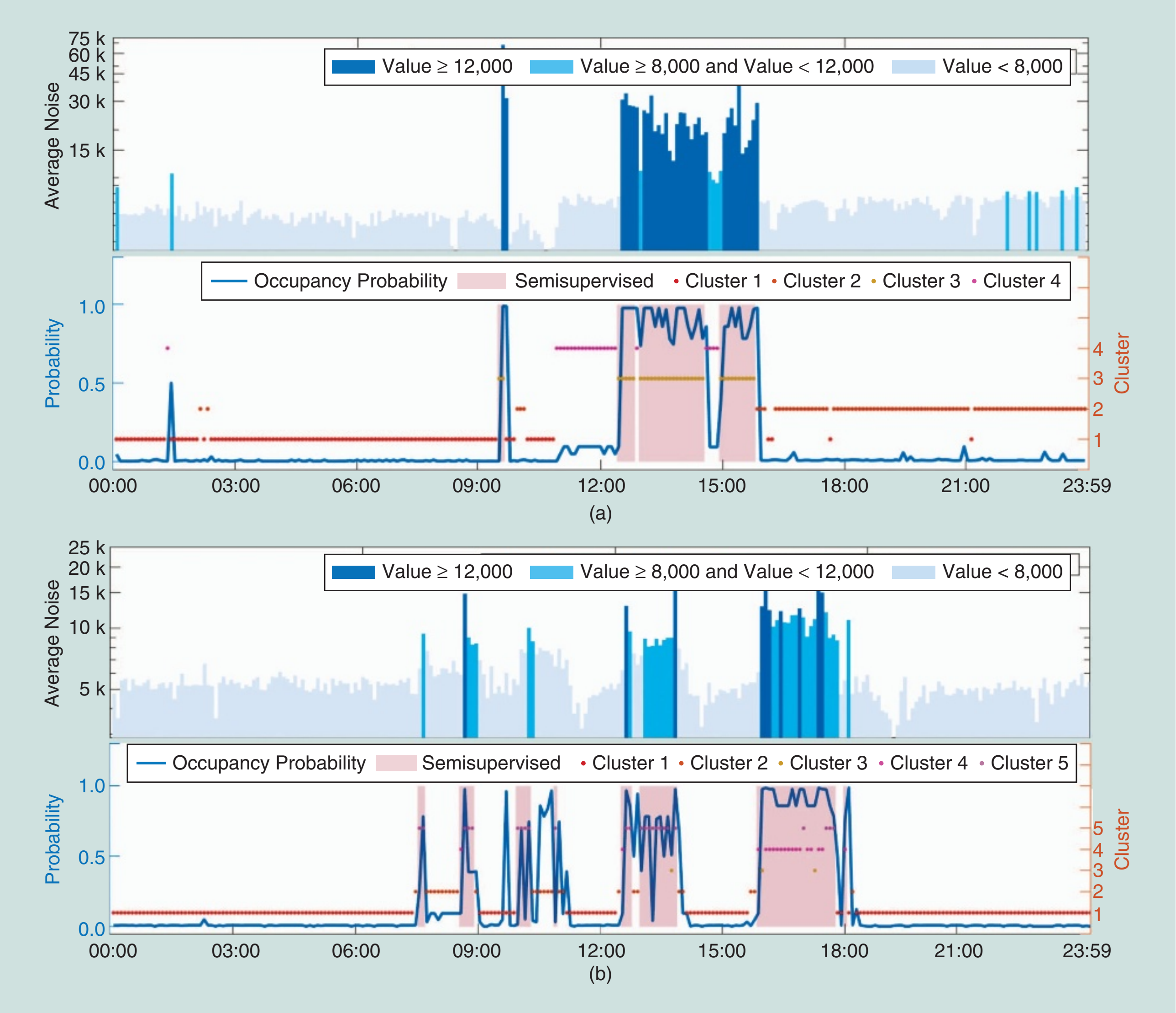


FIGURE 3. The results of occupancy detection in two different meeting rooms (labeled P02 and P04). We visited the site for one day and recorded the ground truth for evaluation. (a) Meeting room P02 and (b) meeting room P04.

as inputs for the deep classifier. In the training stage, we use a data set of 5,000, half of which are used to train the autoencoder. In both cases, training is performed in 30,000 training epochs. As mentioned earlier, we use only a single class to train the autoencoder. We can thereby improve the overall classification accuracy by approximately 3% compared to training using both classes.

Semisupervised learning method

To improve detection accuracy, we use the following semisupervised learning technique. For a particular cluster based on unsupervised learning, we consider the corresponding probability obtained by the deep classifier. If the majority say "occupied," then we label that whole cluster "occupied"; if the majority say "unoccupied," we label that whole cluster "unoccupied." Therefore, we obtain more consistent results compared to using the deep classifier only.

The results for these methods are depicted in Figure 3 for two different meeting rooms, P02 and P04. To summarize, we can observe that the optimal threshold values are different for different rooms (e.g., 12,000 for P02 and 800 for P04): finding those values in a large-scale deployment could be tedious. In the following, we demonstrate how semisupervised learning that combines both clustering and a deep classifier overcomes this problem. We train the deep classifier using data from another two meeting rooms (P01 and P06, with thresholds of 11,000 and 12,000, respectively), which we verify on P02 and P04, as shown in Figure 3. Finally, we see that the semisupervised learning method provides robust and consistent occupancy detection for P02 and P04, even though the training set comes from P01 and P06.

Building efficiency via IoT-based BMSs

In this section, we demonstrate how our designed IoT-based BMS can provide insights for buildings' HVAC efficiency in the future. The HVAC system, in particular, is chosen because this system is responsible for more than 50% of the energy consumption in commercial buildings. A commercial facility in Singapore is considered as the green building test bed. In the selected building, the HVAC system is set up with a modern chiller plant and central air-handling unit (AHU) that can be managed by a typical BMS. The AHU contains an available-speed supply fan, cooling coils, filters, a mixing box, a return air fan, dampers, and several variable air volume (VAV) terminal units that supply chilled air from the AHU to the terminal zones.

We select a specific area of the commercial facility for experiment, i.e., the meeting room P03 in Figure 4 (which shows the schematic of the selected floor plan). On this floor, there is one AHU as well as several rooms with VAV units. Each room is equipped with IoT sensors and, at the entrances, head count cameras. The multipurpose node is responsible for collecting the surrounding environmental information. The environmental information collected for this experiment includes temperature, humidity, light intensity, motion, and noise. Head count cameras are used to count people who enter and exit the rooms within the selected area.

Note that the occupancy within a selected space has a direct influence on the energy consumption of the HVAC system, which (in conjunction with information on the respective energy consumption pattern of the HVAC system) can be exploited to regulate the energy consumption of the HVAC. For example, in [28], the authors propose a data-driven approach

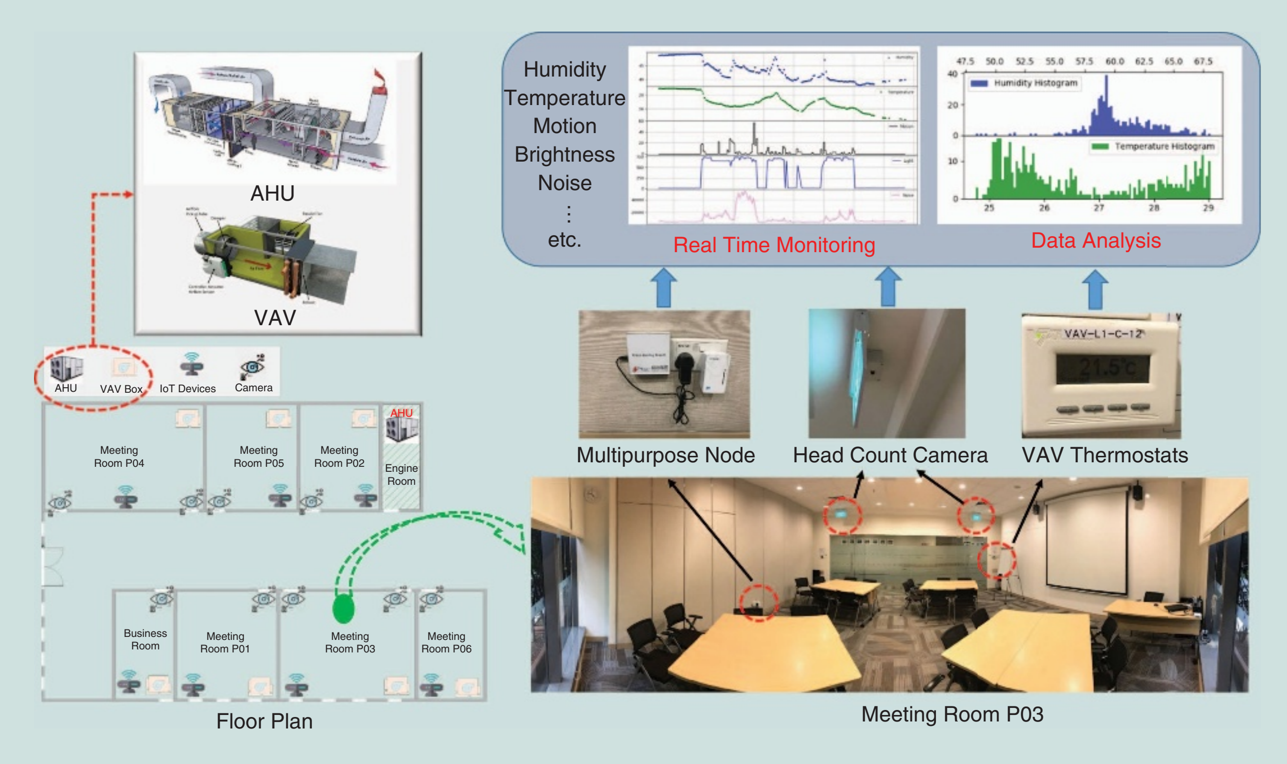


FIGURE 4. A demonstration of the green building test bed.

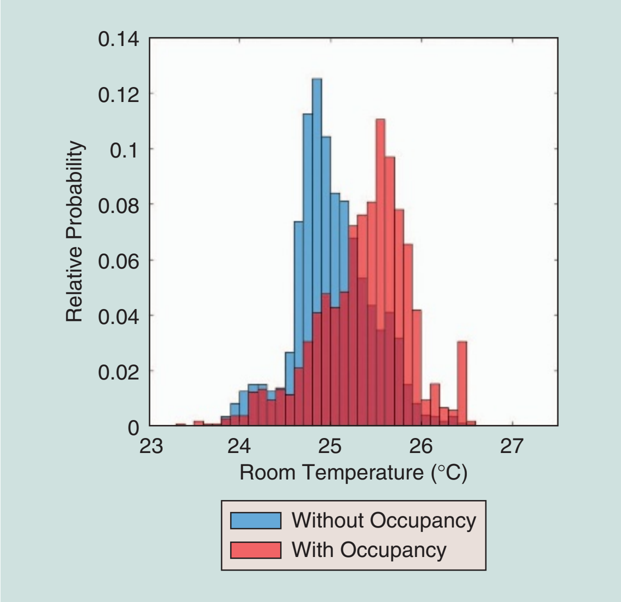


FIGURE 5. An illustration of the distribution of room temperature with and without human occupancy.

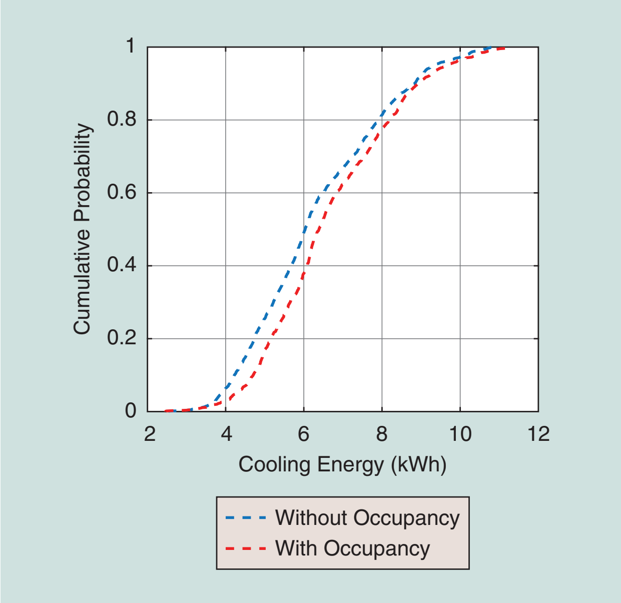


FIGURE 6. A comparison of cooling energy use with and without human occupancy.

for the energy consumption of an HVAC system in which the set-point temperature of the HVAC is regulated according to the people activity and occupancy to control energy consumption. Such control is shown to be very effective and can reduce an HVAC's energy consumption in a house by 33%. An example of another study that uses people occupancy patterns to control the energy consumption of the HVAC can be found in [12].

We select one meeting room of the test bed to demonstrate the energy use analysis. To do so, we extract historical data for the time span between 1 November and 7 December 2017. After filtering out the weekend days and days with missing data (due to maintenance and/or network faults), we have data for a total of 21 days to analyze. The AHU operates from 7:40 a.m. to 7:40 p.m. each weekday; this schedule is maintained by the building manager. However, we consider the office hours from 8:00 a.m. to 7:00 p.m. for our study, because the building is, for the most part, occupied by people only during this period of time.

Based on the collected data on room temperature, chilled air temperature, and air flow, we compute the cooling energy for the energy use analysis. In this context, the required cooling energy for a room is considered as $Q_{\text{cooling}} = \rho_{\text{air}} C_{\text{air}} m_a (T_{\text{room}} - T_{\text{supply}})$ [29], where ρ_{air}, C_{air}, m_a, T_{room}, and T_{supply} represent the air density at 20 °C, the specific heat of air, the volume of air supply, room temperature, and chilled air temperature, respectively. According to the Q_{cooling} formula, the energy consumption of a room for cooling relies on three variables: room temperature, chilled air temperature, and air flow. Among these, chilled air temperature and airflow are controlled by the AHU and are responsible for reducing room temperature to achieve the set point, which is established as 20 °C for the selected room. As such, the higher the room temperature, the more cooling energy is consumed.

The air density and specific heat coefficient are given as 1.204 kg/m^3 and 1.012 kJ/°C/kg, respectively. Then, the room temperature data are extracted for our IoT sensors. The supplied air volume and temperature are monitored by the BMS; hence, the date of volume and temperature can be fetched from the BMS database. We divide the total time period of each experiment into multiple time slots, with each time slot constituting a 5-min interval. Then, we classify the time period with occupancy and without occupancy, respectively, using the occupancy detection technique explained earlier in this section. We obtain two sets of room temperatures for time periods with and without human occupancy. Figure 5 shows the distribution of these two sets of room temperatures. Noticeably, the room temperature is higher when the room is occupied compared to the case without occupancy. Consequently, we infer that more cooling energy will be necessary for consumption during those time periods with occupancy than those with no occupancy.

In addition, we classify the cooling energy usage based on the occupancy status. In Figure 6, we show the cumulative distributions of energy use with and without human occupancy; a gap between the cooling energy use for the two considered occupancy states is clearly visible. Based on this figure, the average cooling energy consumption with occupancy and without occupancy in each time slot is 6.57 and 6.04 kWh, respectively. Hence, the average cooling energy consumption with occupancy is 8.7% higher than that without occupancy. Furthermore, the average room temperature with and without occupancy is 25.37 °C and 25.02 °C, respectively, which indicates an increase of 0.35 °C in average room temperature due to human occupancy.

Based on the previous analysis, it is clear that energy consumed by a building's HVAC system depends to a large extent on its occupancy status, in addition to other factors such as the building's thermal characteristics, use of various appliances, the status of window blinds, and the outdoor climate. Thus, this study provides useful insights concerning the energy efficiency of a building's HVAC systems by exploiting its human occupancy status, which can be determined easily by simply deploying IoT devices in the building. For instance, the occupancy information can be used to develop suitable control strategies or optimization tools via signal processing techniques to opportunistically reduce HVAC energy consumption and subsequently help the occupants reduce their energy costs. For more information about how occupancy information can be used to develop suitable control strategies or optimization tools, readers may refer to [7], [13], and [30].

Conclusions

In this article, we first reviewed the application of IoT-based signal processing techniques for managing various subsystems within a building. Then, we provided an overview of how machine learning can be applied with an IoT device to detect human occupancy within a building, which contributes significantly to the building's energy consumption. In particular, we considered

- a transfer learning-based technique that counts people based on images captured at the entrances of the selected areas of a building
- an unsupervised learning technique labeled by deep-learning-based occupancy detection that uses information obtained by sound sensors.

Further, we provided a short description of the test bed where these techniques are deployed for occupancy detection and showed how the information can help gain insights on the use of the HVAC system within the test bed. The correlation between occupancy and energy use, as demonstrated in this study, has the potential to be used in developing different energy-management schemes that will help reducing energy consumption and electricity costs for building managers.

There are many areas into which the work reported here can be extended.

- *Widespread deployment*: The discussed study was conducted on a relatively small scale in a commercial building in Singapore. While using a real facility provides actual user and system data for our study and analysis, it would be interesting to see how the findings from the study can be extended to a larger scale, e.g., the entire building, via widespread deployment of IoT devices.
- *Detailed modeling of energy usage*: Another interesting extension of the proposed work would be to use machine-learning techniques to perform more detailed modeling of energy use by various appliances in the building and then design suitable techniques to optimize this use to reduce electricity costs.
- *Applying the IoT beyond building energy*: IoT sensors also provide valuable data for predictive maintenance and anomaly detection. Hence, it would be interesting to explore how the collected data from IoT sensors can be used to help a building perform asset management, which, in addition to energy reduction, may reduce other building costs.
- *Exploring quantitative performance*: This article presents qualitative results for the head counting and occupation detection study and their importance for building management. However, there is a need to extend this work to present more quantitative analysis in terms of performance compared with existing studies in the literature and analyze how head counting impacts the energy consumption of the building.

Acknowledgments

This work was supported, in part, by project NRF2015ENC-GBICRD001-028 funded by the National Research Foundation (NRF) of Singapore via the Green Buildings Innovation Cluster (GBIC) administered by the Building and Construction Authority (BCA)–Green Building Innovation Cluster (GBIC) Program Office; in part, by the SUTD-MIT International Design Center (IDC; idc.sutd.edu.sg); in part, by the grant NSFC 61750110529; and, in part, by the U.S. National Science Foundation under grants CNS-1702808 and ECCS-1549881. Any findings, conclusions, or opinions expressed in this document are those of the authors and do not necessarily reflect the views of the sponsors.

Authors

Wayes Tushar (wayes.tushar.t@ieee.org) received the B.Sc. degree in electrical and electronic engineering from the Bangladesh University of Engineering and Technology in 2007 and the Ph.D. degree in engineering from the Australian National University in 2013. Currently, he is working as an Advance Queensland research fellow at the School of Information Technology and Electrical Engineering, University of Queensland. His research interests include energy and storage management, renewable energy, smart grid, design thinking, and game theory.

Nipun Wijerathne (hnipun@gmail.com) received the B.Sc. degree from the Department of Electronic and Telecommunication Engineering, University of Moratuwa, Sri Lanka, in 2016. Then, he joined the Singapore University of Technology and Design as a researcher. His research interests include machine learning, deep learning, signal processing, and Bayesian learning methods for practical problems. He is currently working as a machine-learning engineer at Cloud Solutions International, Sri Lanka.

Wen-Tai Li (wentai_li@sutd.edu.sg) received the B.S. and M.S. degrees in optoelectronics and communication engineering from National Kaohsiung Normal University, Taiwan, in 2009 and 2011, respectively, and the Ph.D. degree in communications engineering from National Sun Yat-sen University, Kaohsiung, Taiwan, in 2018. He was a research assistant from 2015 to 2017 and is currently a postdoctoral fellow with the Singapore University of Technology and Design. His research interests include power system monitoring, smart grid, and wireless communications.

Chau Yuen (yuenchau@sutd.edu.ag) received the B.Eng. and Ph.D. degrees from Nanyang Technological University, Singapore, in 2000 and 2004, respectively. He was a postdoctoral fellow with Lucent Technologies–Bell Labs, Murray Hill, New Jersey, in 2005. He was a visiting assistant professor with Hong Kong Polytechnic University in 2008. From 2006 to 2010, he worked at I2R, Singapore, as a senior research engineer. Currently, he is an associate professor with the Singapore University of Technology and Design. He is an editor of *IEEE Transactions on Communications* and *IEEE Transactions on Vehicular Technology*. In 2012, he received the IEEE Asia-Pacific Outstanding Young Researcher Award.

H. Vincent Poor (poor@princeton.edu) is the Michael Henry Strater University Professor of Electrical Engineering at Princeton University, New Jersey. His interests include information theory and signal processing, with applications in wireless networks, energy systems, and related fields. He is an IEEE Fellow, a member of the National Academy of Engineering and National Academy of Sciences, and a foreign member of the Chinese Academy of Sciences and the Royal Society. He received the IEEE Signal Processing Society Technical Achievement and Society Awards in 2007 and 2011, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal and a D.Sc. *honoris causa* from Syracuse University, also in 2017.

Tapan Kumar Saha (saha@itee.uq.edu.au) received his B.Sc. degree in electrical and electronics engineering from the Bangladesh University of Engineering and Technology, Dhaka, in 1982; M.Tech. degree in electrical engineering from the Indian Institute of Technology Delhi, in 1985; and Ph.D. degree from the University of Queensland, Brisbane, Australia, in 1994. He is currently a professor of electrical engineering with the School of Information Technology and Electrical Engineering, University of Queensland. His research interests include condition monitoring of electrical assets, power systems, and power quality. He is a fellow of the Institution of Engineers, Australia.

Kristin L. Wood (kristinwood@sutd.edu.sg) received his M.S. and Ph.D. degrees in the Division of Engineering and Applied Science at the California Institute of Technology, Pasadena. He is currently a professor with the Singapore University of Technology (SUTD) as well as founding head of Pillar, Engineering, and Product Development; associate provost for graduate studies; and codirector of the SUTD-MIT International Design Center. He has published more than 450 refereed articles and books; received more than 80 awards in design, research, and education; and consulted with more than 100 companies worldwide. He is a fellow of the American Society of Mechanical Engineers.

References

[1] W. Tushar, C. Yuen, W.-T. Li, D. Smith, T. Saha, and K. L. Wood, “Motivational psychology driven AC management scheme: A responsive design approach,” *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 1, pp. 289–301, Mar. 2018.

[2] M. Manic, D. Wijayasekara, K. Amarasinghe, and J. J. Rodriguez-Andina, “Building energy management systems: The age of intelligent and adaptive buildings,” *IEEE Ind. Electron. Mag.*, vol. 10, no. 1, pp. 25–39, Mar. 2016.

[3] J. Pan, R. Jain, S. Paul, T. Vu, A. Saifullah, and M. Sha, “An Internet of Things framework for smart energy in buildings: Designs, prototype, and experiments,” *IEEE Internet Things J.*, vol. 2, no. 6, pp. 527–537, Dec. 2015.

[4] A. Pandharipande and D. Caicedo, “Smart indoor lighting systems with luminaire-based sensing: A review of lighting control approaches,” *Energy Buildings*, vol. 104, pp. 369–377, Oct. 2015.

[5] Z. Cheng, Q. Zhao, F. Wang, Y. Jiang, L. Xia, and J. Ding, “Satisfaction based Q-learning for integrated lighting and blind control,” *Energy Buildings*, vol. 127, pp. 43–55, Sept. 2016.

[6] S. Jain and V. Garg, “A review of open loop control strategies for shades, blinds and integrated lighting by use of real-time daylight prediction methods,” *Building Environ.*, vol. 135, pp. 352–364, May 2018.

[7] X. Guo, D. Tiller, G. Henze, and C. Waters, “The performance of occupancy-based lighting control systems: A review,” *Lighting Res. Technol.*, vol. 42, no. 4, pp. 415–431, Aug. 2010.

[8] B. Sun, P. B. Luh, Q. S. Jia, Z. O’Neill, and F. Song, “Building energy doctors: An SPC and Kalman filter-based method for system-level fault detection in HVAC systems,” *IEEE Trans. Autom. Sci. Eng. (from July 2004)*, vol. 11, no. 1, pp. 215–229, Jan. 2014.

[9] S. Ali and D.-H. Kim, “Effective and comfortable power control model using Kalman filter for building energy management,” *Wirel. Personal Commun.*, vol. 73, no. 4, pp. 1439–1453, Dec. 2013.

[10] A. Javed, H. Larijani, A. Ahmadinia, and D. Gibson, “Smart random neural network controller for HVAC using cloud computing technology,” *IEEE Trans. Ind. Informat.*, vol. 13, no. 1, pp. 351–360, Feb. 2017.

[11] A. Al-Ali, I. A. Zualkernan, M. Rashid, R. Gupta, and M. AliKarar, “A smart home energy management system using IoT and big data analytics approach,” *IEEE Trans. Consum. Electron.*, vol. 63, no. 4, pp. 426–434, Nov. 2017.

[12] M. Aftab, C. Chen, C.-K. Chau, and T. Rahwan, “Automatic HVAC control with real-time occupancy recognition and simulation-guided model predictive control in low-cost embedded system,” *Energy Buildings*, vol. 154, pp. 141–156, Nov. 2017.

[13] A. Mirakhorli and B. Dong, “Occupancy behavior based model predictive control for building indoor climate: A critical review,” *Energy Buildings*, vol. 129, pp. 499–513, Oct. 2016.

[14] D. Zhang, S. Li, M. Sun, and Z. O’Neill, “An optimal and learning-based demand response and home energy management system,” *IEEE Trans. Smart Grid*, vol. 7, no. 4, pp. 1790–1801, July 2016.

[15] S. Singh and A. Majumdar, “Deep sparse coding for non-intrusive load monitoring,” *IEEE Trans. Smart Grid*, 2017. doi: 10.1109/TSG.2017.2666220.

[16] W. Kong, Z. Y. Dong, D. J. Hill, F. Luo, and Y. Xu, “Short-term residential load forecasting based on resident behaviour learning,” *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 1087–1088, Jan. 2018.

[17] A. Belmonte-Hernández, G. Hernández-Peñaloza, F. Álvarez, and G. Conti, “Adaptive fingerprinting in multi-sensor fusion for accurate indoor tracking,” *IEEE Sensors J.*, vol. 17, no. 15, pp. 4983–4998, Aug. 2017.

[18] W. Li, Y. Lu, J. Sun, Q. Chen, T. Dong, L. Zhou, Q. Zhang, and L. Wei, “People counting based on improved gauss process regression,” in *Proc. Int. Conf. Security, Pattern Analysis and Cybernetics*, Shenzhen, China, Dec. 2017, pp. 603–608.

[19] A. Tyndall, R. Cardell-Oliver, and A. Keating, “Occupancy estimation using a low-pixel count thermal imager,” *IEEE Sensors J.*, vol. 16, no. 10, pp. 3784–3791, May 2016.

[20] Y. P. Raykov, E. Ozer, G. Dasika, A. Boukouvalas, and M. A. Little, “Predicting room occupancy with a single passive infrared (PIR) sensor through behavior extraction,” in *Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Computing*, Heidelberg, Germany, Sept. 2016, pp. 1016–1027.

[21] J. Salamon and J. P. Bello, “Unsupervised feature learning for urban sound classification,” in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing*, Brisbane, Australia, Apr. 2015, pp. 171–175.

[22] S. K. Viswanath, C. Yuen, W. Tushar, W.-T. Li, C.-K. Wen, K. Hu, C. Chen, and X. Liu, “System design of the Internet of Things for residential smart grid,” *IEEE Wireless Commun.*, vol. 23, no. 5, pp. 90–98, Oct. 2016.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in *Proc. European Conf. Computer Vision*, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.

[24] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, “Speed/accuracy trade-offs for modern convolutional object detectors,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, July 2017, pp. 3296–3297.

[25] Y. Jeon, C. Cho, J. Seo, K. Kwon, H. Park, S. Oh, and I.-J. Chung, “IoT-based occupancy detection system in indoor residential environment,” *Buildings Environ.*, vol. 132, pp. 181–204, Mar. 2018.

[26] L. M. Candanedo and V. Feldheim, “Accurate occupancy detection of an office room from light, temperature, humidity, and CO_2 measurements using statistical models,” *Energy Buildings*, vol. 112, pp. 28–39, Jan. 2016.

[27] L. P. L. Billy, N. Wijerathne, B. K. K. Ng, and C. Yuen, “Sensor fusion for public space utilization monitoring in a smart city,” *IEEE Internet Things J.*, vol. PP, no. 99, pp. 1–1, 2017.

[28] W. T. Li, S. R. Gubba, W. Tushar, C. Yuen, N. U. Hassan, H. V. Poor, K. L. Wood, C, and K. Wen, “Data driven electricity management for residential air conditioning systems: An experimental approach,” *IEEE Trans. Emerg. Topics Comput.*, 2017. doi: 10.1109/TETC.2017.2655362.

[29] B. Balaji, H. Teraoka, R. Gupta, and Y. Agarwal, “ZonePAC: Zonal power estimation and control via HVAC metering and occupant feedback,” in *Proc. ACM Workshop Embedded Systems Energy-Efficient Buildings*, Rome, Italy, Nov. 2013, pp. 18:1–18:8.

[30] J. Yang, M. Santamouris, and S. E. Lee, “Review of occupancy sensing systems and occupancy modeling methodologies for the application in institutional buildings,” *Energy Buildings*, vol. 121, pp. 344–349, June 2016.

Viswam Nathan, Sudip Paul, Temiloluwa Prioleau,
Li Niu, Bobak J. Mortazavi, Stephen A. Cambone,
Ashok Veeraraghavan, Ashutosh Sabharwal,
and Roozbeh Jafari

A Survey on Smart Homes for Aging in Place

Toward solutions to the specific needs of the elderly

INTERNET OF THINGS—ISTOCKPHOTO.COM/IAREMENKO
CIRCUITS—IMAGE LICENSED BY INGRAM PUBLISHING

Advances in engineering and health science have brought a significant improvement in health care and increased life expectancy. As a result, there has been a substantial growth in the number of older adults around the globe, and that number is rising. According to a United Nations report, between 2015 and 2030, the number of adults over the age of 60 is projected to grow by 56%, with the total reaching nearly 2.1 billion by the year 2050 [1]. Because of this, the cost of traditional health care continues to grow proportionally. Additionally, a significant portion of the elderly have multiple, simultaneous chronic conditions and require specialized geriatric care. However, the required number of geriatricians to provide essential care for the existing population is four times lower than the actual number of practitioners, and the demand-supply gap continues to grow [2]. All of these factors have created new challenges in providing suitable and affordable care for the elderly to live independently, more commonly known as *aging in place*.

Prioritizing proactive care

The driving goal is to keep individuals, especially the elderly, healthy at home through proactive care, while also facilitating remote reactive care when needed rather than requiring frequent visits to the doctor. Proactive care can minimize the physical and mental stress associated with regular hospitalization for the elderly and significantly reduce the financial burden for both patients and the health-care system. Both proactive and remote reactive care can be enabled through continuous, holistic monitoring of the user's health status, daily activities, and behavioral patterns with multimodal sensors in a naturalistic environment. The simultaneous application of both proactive and reactive care, in turn, can promote the use of diagnostic testing to avoid adverse medical outcomes and alleviate the burden on the user as well as the health-care system. In this regard, homes equipped with sensors and smart systems, also known as *smart homes*, designed for the benefit of the aging residents will enable both short-term monitoring for remote reactive care (e.g., monitoring cardiac activity in

Digital Object Identifier 10.1109/MSP.2018.2846286
Date of publication: 28 August 2018

1053-5888/18©2018IEEE

response to newly prescribed medication) and long-term monitoring for proactive care (e.g., tracking adherence to prescribed exercise routines or suggesting lifestyle modifications based on observed behavioral trends of the user).

By using wearable and environmental sensors, wireless sensor networks, and sensing devices that can monitor critical health parameters, we will be able to gather physiological and behavioral data continuously. The key lies in building intelligent, efficient algorithms that can provide valuable insights from daily patterns, e.g., from the changes in a user's gait patterns or eating habits. The new algorithms could also enable predictions of future irregularities, allowing us to turn these predictions into actionable information impacting the quality of life and care delivery, while offering opportunities for adaptive interventions and personalized medicine.

In this article, we present several noteworthy investigations in the field of smart homes and assisted living applications and categorize the important research areas. In particular, we discuss the required technology support including 1) sensors and connectivity solutions, 2) signal processing and data analytics that operate on sensor data and extract actionable information, and 3) information delivery and visualization paradigms that provide the actionable information to the end-users and stakeholders. We will provide in-depth analysis of the challenges and discuss the benefits of smart home technology for aging in place.

While there have been some previous surveys done on smart homes, these surveys are either limited in their scope or fail to provide research direction or opportunities. For example, one recent survey focuses on smart home technologies for activity recognition and its impact on health care in general [3], while another focused on the technology readiness and the effectiveness of existing smart home technologies to address some of the complex health issues faced by older adults but does not provide any guidelines for future researchers [4]. Unlike these previous surveys, this work focuses on the key challenges associated with aging in place for the elderly, considers diverse application drivers, and presents a clear outline of the most important research areas and opportunities for future work.

Application case studies: Major smart home projects for aging in place

In general, smart home projects for aging in place mainly focus on monitoring the health and well-being of elderly persons through sensors tagged on the habitat (doors, walls, ceiling, and so on), sensors in/on household objects (small appliances, beds, couches, and others), or sensors worn by the users (i.e., wearable devices). Monitoring the daily activities and behavioral patterns of a dweller in a smart home environment can be a key factor for facilitating aging in place. Daily activities can include basic efforts like walking, sleeping, personal grooming, toilet usage, self feeding, and so on or instrumental tasks such as food preparation, cleaning, the use of communication tools (e.g., telephones), and watching TV, along with others. Professionals in health care use the term *activities of daily living* (*ADL*) to refer to the daily activities that a person normally performs, and it can be used to define functional capacity, especially that of an elderly person.

Several studies have investigated systems based on habitat sensor networks and household object sensor networks for monitoring ADL. Suryadevara et al. propose an activity detection system consisting of a wireless sensor network, which has the objective of wellness detection [5]. The sensor network consists of current sensors placed at power outlets to detect the use of electrical appliances, flexible pressure sensors to detect activity around nonelectrical objects (e.g., beds and sofas), and a Zigbee-based mesh network protocol for connectivity. Through the use of a conditional probabilistic model, this method achieves an overall accuracy of 94% in detecting and forecasting daily activities. Another study that leverages wireless sensor networks for recognizing ADL was conducted by Ghayvat et al. [6]; it also uses a Zigbee mesh topology for the network. Their system employs power outlet sensors for monitoring the use of electrical and electronic devices, pressure and contact sensing, passive infrared (PIR)-based movement sensing, and temperature-monitoring units, all connected through Zigbee-based radio-frequency (RF) modules. This particular investigation also reports on the interference and attenuation issues of the wireless network when implemented in a smart home. A common factor between these two investigations, besides ADL recognition, is the use of parameters called *wellness indices* or *wellness functions* to quantify the well-being of an elderly person. Both studies leverage the measurement of active versus inactive time intervals of different appliances to estimate wellness.

Fleury et al. propose a support vector machine (SVM)-based ADL recognition mechanism leveraging a variety of sensors including wide-angle webcams, microphones, and contact sensors to detect the opening and closing of doors, infrared (IR) sensors to detect the presence of a subject in a room, and wearable motion sensors (i.e., three-axis accelerometer and magnetometer) [7]. The authors suggest tagging the habitat with sensors rather than objects to simplify the design and implementation. This work considers seven basic ADL (including eating, sleeping, toilet usage, and so on) and uses multiclass SVMs to classify them, with accuracies ranging from 97.8% to 64.3% depending on the activity.

While the prior studies reported impressive accuracy levels for detecting ADL, these systems have been designed with static requirements and do not consider the possible variations in the application and usage of the system and variations in the environment. The detection and system architecture requirements may vary from one application to another, and it is important to maintain the concept of adaptability and tuning of the accuracy, sensitivity and the specificity requirements of the ADL detection. This typically translates to tuning of system architecture, sensors, and the optimization process, accordingly. Additionally, the output of the ADL recognition may need to be represented in various forms, including deterministic and probabilistic, which is not present in the proposed case studies.

Moving away from detection of ADL, Kim et al. proposed an alternative method of inferring the well-being of an elderly person using location information [8]. Their method incorporated an RF identification (RFID)-based indoor tracking system that used location information in association with the durations of stay in different locations of the home to infer information such as movement patterns, and the frequency of certain location-specific activities (e.g., using the toilet or sleeping in the bedroom) to estimate the well-being of an elderly dweller.

While most investigations in smart home technology consider only a single user, this is not necessarily the case in a real home environment. Moreover, recognizing multiple users could enable detection of social interactions associated with wellness. The recognition of ADL in a multiuser setting is challenging due to two additional requirements: 1) identifying each of the dwellers and 2) accommodating a more complex set of activities involving multiple persons. Wang et al. presented a multiuser activity recognition system in a smart home setting using wearable audio sensors, actimetry sensors (e.g., accelerometer, temperature, humidity, and light), and RFID tags [9]. This system achieved a maximum accuracy of 98.59% in detecting single-user activities and 95.91% in detecting multiuser activities. While these results are admirable, one shortcoming is that the single- and multiuser activities are predefined and distinct; this may not be the case in a real-world scenario where these two different types of activities can overlap. Another system by Mokhtari et al., which uses PIR-based occupancy sensors and ultrasound arrays, performs human identification among multiple users and reports 100% accuracy [10]. This system, which uses Bluetooth Low Energy for connectivity and has been designed with energy efficiency in mind, recognizes different users based on their height and detects movement direction and speed to monitor a user. One shortcoming of this system is that the height difference between each of the users has to be at least 4 cm for the algorithms to operate with an acceptable level of accuracy.

An open question is a uniform, generalizable, and quantifiable description of the well-being of an older adult. While two of the studies mentioned in this section presented wellness indices, they each had different definitions for the term; the research community for this application space could benefit from a more standardized definition of this wellness index to appropriately assess the effectiveness of smart home systems in estimating wellness. A standardized definition can also help determine ADL of relevance, which in turn can dictate the number and type of sensors used in the smart home. One approach that has been previously explored to bridge this gap is to establish relationships between recognized ADL and clinically established mobility and cognitive tests such as Timed Up and Go and Repeatable Battery for the Assessment of Neuropsychological Status [11]. Additionally, there are opportunities to create such generalizable or disorder-specific wellness indices, potentially customized to individuals by comparing the observed trends to each user's baselines, thus offering insights for improvement and progress.

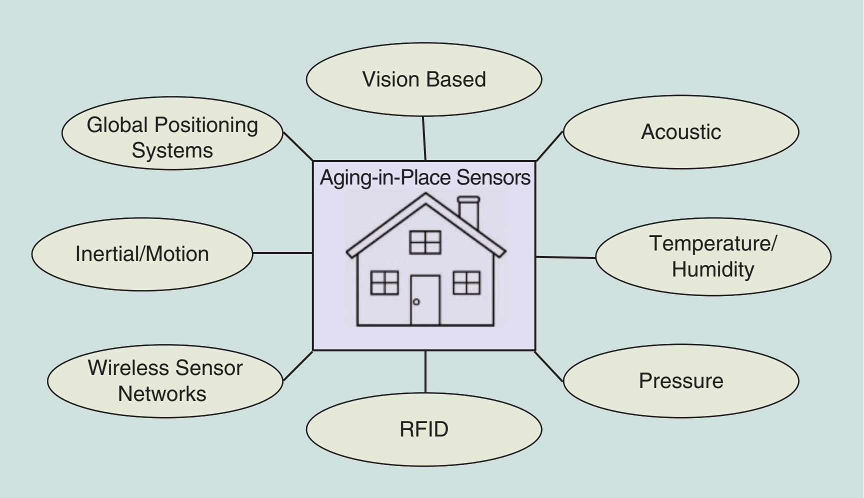


FIGURE 1. The common sensors that are used to support aging in place.

Sensors and connectivity paradigms

Sensors are crucial for measuring data from individuals and environments. These sensors can be discrete (e.g., contact switches) or continuous (e.g., physiological sensors) observing devices [3]. Figure 1 presents an overview of various sensor types that have been used with the diverse range of complex monitoring and automation tasks for aging in place.

It is rarely the case that a single sensor type is sufficient for quantifying the health and well-being of a person; therefore, multiple sensors are often combined to achieve specific goals. For the same target phenomenon, different sensors have their own observations with different levels of noise and reliability. In a survey on fall detection and activity recognition in elderly care by Abbate et al. [12], the authors highlight the use of vision-based sensors and environmental and/or wearable sensors such as inertial sensors for fall detection in a home setting. Vision-based sensors are reliable and can depend on sophisticated image-recognition algorithms; however, it is costly and time consuming to install these cameras, and there are significant privacy concerns associated with this modality. Environmental sensors, such as IR or pressure sensors placed on household objects, offer a cheaper alternative that preserves privacy; yet they are limited to sensing only the specific spaces/objects on which they are placed. Wearable sensors, such as an inertial sensor on an ankle strap, are user centric and allow ubiquitous, unrestricted monitoring at low cost, unlike vision and environmental sensors. Nevertheless, data from wearable sensors can be challenging to interpret due to noise from motion artifacts, misleading data due to improperly worn sensors, or missing data due to sensors occasionally not being worn at all. There is an opportunity to develop generalizable sensor selection techniques that consider the complementary nature of the sensors in the context of the end-application requirement.

Functional monitoring is also particularly important in smart homes to support aging in place for the elderly. Research supports the notion that a variety of factors, including physical and intellectual activity, social engagement, and nutrition, all contribute to optimizing cognitive health in the aging population [13]. To enable the monitoring of mental health, a combination of different sensing modalities is imperative, such as using PIR or inertial sensors for physical activity monitoring, acoustic sensors for social monitoring, and vision-based sensors for nutrition assessment. In addition to functional monitoring, physiological monitoring is also of particular importance toward achieving the goal of aging in place. This can include the detection of emergency situations such as falls using wireless networks [12], continuous monitoring of existing chronic conditions, e.g., dementia or cardiac health [14], and monitoring of sleep health using motion sensors [15]. Wood et al. presented a wireless sensor network system called *AlarmNet*, which integrates environmental, physiological, and activity sensors in a single architecture [16]. The AlarmNet system is unique because it enables improved power conservation by anticipating which sensors should be active and which should be disabled by analyzing the behavioral pattern of the user. Additionally, the system is designed with flexibility, which allows for the integration of new sensors and ad hoc deployment into existing structures.

Sensor selection and ease of deployment are a critical challenges in the design of smart homes for aging in place. The types and number of sensors to be deployed should not become a burden for the user. Human factors, such as ease of use even with declining levels of function and cognition, must be taken into consideration when designing and deploying the sensors [17]. Some key factors that contribute to technology acceptance, particularly among older adults, include perceived usefulness and ease of use, as well as personal characteristics, e.g., functional ability [18]. In noncritical cases, it is unlikely that older adults will be inclined to keep up with constantly changing technology developments in the form of new wearables and environmental technology to be deployed in the home. Therefore, there is an opportunity to leverage existing sensors that were designed for a different purpose for a new sensing paradigm that can, for example, evaluate mobility, social engagement, and loneliness [19]. Additionally, minimizing the number of sensors required would ease communication bandwidth, energy efficiency, costs, and user acceptance concerns. For example, wireless sensor networks can be used not only for daily activity monitoring but also for monitoring sociability and detecting emergencies. To facilitate this, sensors provide the recorded data as well as a quality measure for the data, e.g., a wrist-worn heart-rate monitor can not only detect the heart rate but also the confidence in the heart rate observations and the quality associated with the data, which could be impacted by motion artifacts and can be measured by motion sensors.

The connectivity among different sensors can be realized using wireless sensor networks comprising Bluetooth, RF, Zigbee, RFID, ONE-NET, Wi-Fi and so on and even wired connections like serial communication, Multimedia over Coax Alliance, and Ethernet to create a smart environment. However, challenges exist with numerous communication protocols that are often incompatible at various networking layers, and the existence of varying throughput requirements on sensor outputs and data increases communication complexity [20]. Communication among several high- and low-end sensor nodes has to be established while taking into consideration constraints of the sensor nodes in question. There is currently no unified software interface to collect data from sensors, and this makes it challenging to interface existing sensor data streams with new software since the requirements and specifications are often incompatible. This presents an opportunity for the development of a unified and standard software interface for sensors.

With the growing number of sensors and interconnected systems for aging in place, the requirements for privacy of personal data and secured end-to-end connections become critical. Security can be enforced on two levels: device-level security includes hardware encryption and access control in stand-alone devices, while network- and system-level security include encryption of network traffic, source blocking, and authentication.

Smart home technologies use a diverse range of communication techniques, and the use of Internet protocol connectivity provides the bridge among these devices [21]. However, the use of Internet communication brings the challenge of dealing with cybertheft, data manipulation, unauthorized access, and other such undesirable events. Organizations like the Internet Engineering Task Force continue to work toward the standardization of security in data exchange protocols and enhancing Datagram Transport Layer Security [21]. One investigation of network security by Sivaraman et al. proposes augmentation of network-level security measurements with device-level protection and implements a prototype consisting of a third-party architecture and associated application programming interfaces [22]. The authors also report the security vulnerability of some commercial Internet of Things products and evaluate their implemented software-defined network platforms' protection efficacy.

A number of cryptography methods are used in a variety of security scenarios. RFID-based authorization schemes are seeing increased use, and elliptic curve cryptography is a popular technique used in health-care environments. In their review of several recent works on elliptic-curve-cryptography-based RFID schemes, He et al. considered computation and communication costs as well as several security requirements to compare performances [23]. The authors report that very few works satisfy all security requirements while keeping the cost in an acceptable range. Thus, the establishment of secure protocols for communication among devices subject to the application requirements remains an important research opportunity to realize smart homes for aging in place.

Signal processing and data analytics

Signal processing and data analytics in the context of a smart home signify the effective fusion of data from multiple heterogeneous sensors, knowledge extraction, and production of actionable information (see Figure 2). Signal processing is required to process noisy signals to observe fine-grained information over short time periods, such as the response to blood pressure medication over several minutes/hours. In contrast, data analytics can provide more coarse-grained information over several weeks/months after recognizing long-term trends and potentially making predictions and providing feedback to stimulate behavioral changes. Moreover, in a smart home environment with a multitude of sensors tracking the user's location and activity, these approaches can exploit the knowledge of context to improve estimates.

Many different signal processing techniques have been implemented for the purpose of monitoring the health and well-being of occupants in smart homes. Zheng et al. used a self-adaptive neural network called *Growing Self-Organizing Maps* (*GSOM*) for human activity detection in a smart home environment [24]. Starting with an initial network composed of four neurons on a two-dimensional (2-D) grid, the GSOM network adapted during training to determine the winning neuron for each input data and updated the associated weight vectors. One drawback of this approach is that several parameters of the network need to be determined in advance through heuristic trial and error; hence, there is scope here to augment this approach with other machine-learning techniques. Apart from traditional learning methods such as the SVM, some recent machine-learning methods such as temporal neural networks, the hierarchical hidden semi-Markov model, and intertransaction association rule have been used for activity recognition in smart homes and assisted living spaces. Another machine-learning approach is to strategically combine knowledge from various sensors to validate the extracted knowledge and minimize false alarms in emergency detection. Tabar et al. combined wearable accelerometers based on a threshold-based method to detect sudden movements of the user [25]. These sudden events triggered an environmental camera within the space to perform position estimation using simultaneous visual observations and vision-based reasoning. These multisensor learning approaches dovetail well with the requirement for multiple heterogeneous sensors, as described in the previous section.

One challenge is to design algorithms in such a way that the required number of sensors for a given application is optimized [16]. Relying on too few sensors increases the likelihood of the algorithm producing false alarms, which is undesirable, especially in the case of emergency-aware applications. Conversely, an algorithm that relies on too many sensors increases complexity, causing energy and resource consumption for the target smart home system to rise.

Adaptive learning models like GSOM allow the learned algorithms to change and improve with the constantly changing physical environment [24]. For example, a two-occupant home can temporarily become a one-occupant home due to illness, travel, or a change in schedule, and sensors may be added or removed from the network arbitrarily. Therefore, fixed-learning models pose a challenge and can quickly become obsolete. Additionally, the framework should accommodate customizable models for different users; training for specific users will likely focus the accuracy of the learned model on the information that each individual provides, as opposed to expecting a generalized single model to fit a diverse, heterogeneous population. Therefore, an opportunity exists to automatically establish customizable learning models. Given the dynamic nature of the sensor network, transfer learning becomes an important opportunity, i.e., effectively transferring the user behavior and parameters learned through one set of sensors in one environment to another set of sensors in a new environment.

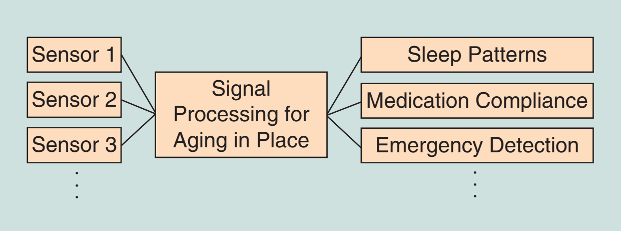


FIGURE 2. The objective of signal processing for aging in place.

Furthermore, large amounts of unlabeled data sets from in-home settings already exist; therefore, a research opportunity lies in improving automatic or semiautomatic labeling techniques of unknown or new data streams. Manual labeling of large volumes of data is not feasible or realistic, so algorithms that can label new data sets from the extracted knowledge gained from a small training set are highly attractive and desired.

A smart home collecting data about a user continuously and persistently via multiple sensors represents a valuable base for longitudinal studies. Evaluations of the performance of certain physical activities can be used to predict health conditions in the older population, while meaningful change detection over time is crucial for proactive health care. Active intervention to help users mold their habits and activities can be the key to proactive health care, as opposed to reactively attending to adverse events after they happen.

However, it is often difficult for users themselves to observe the subtle changes over time because chronic syndromes often progress very slowly and short-term observations may be quite noisy. Thus, an intelligent home-based system is required for continuous, longitudinal, and unobtrusive assessment of the life pattern changes of dwelling older adults. Moreover, by observing the changes over time, an algorithm can predict future trends and possibly avoid undesirable outcomes. In smart home environments, sensor data have been widely used for longitudinal and continuous study concerning daily activities, sleep patterns, gait velocity, and loneliness, among others.

Unusual routines of residents are identified by tracking their mobility and recognizing their daily activities (e.g., sleeping, grooming, and eating) based on the collected sensor data, resulting in an interesting observation on the statistically significant correlation between the changes of daily activities (e.g., mobility scores) and the changes in a clinical measure of global cognitive health [26].

Sleep patterns, a special case of daily activities, convey information critical for assessing human wellness. Mihailidis et al. focused on studying sleep patterns and proposed an approach to measure sleep hygiene of elders over a six-month period [15] in which both acute and slow changes in the sleep patterns are successfully identified. Considering the potential for gait velocity to predict morbidity and mortality, Hagler et al. estimated walking speed from noisy time and location data collected by a sensor line of restricted-view PIR motion detectors [27]. For the approaches presented in both of these works, there is still an open research opportunity to validate the measured trends with known clinical measures to ensure the recorded long-term data are beneficial. Besides the aforementioned unusual physical behavior, mental aspects such as loneliness are also closely related to increased morbidity and mortality, which may lead to decreased sleep quality and increased risk of cognitive decline [19]. Nevertheless, assessing loneliness in older adults is challenging due to the negative desirability bias associated with being lonely. To circumvent this problem, Austin et al. propose a system to measure loneliness by assessing in-home behavior using wireless motion sensors, contact sensors, and phone monitors [19].

One important challenge that these kinds of long-term studies face is the lack of a gold-standard ground truth for the target parameters. It is extremely difficult to track the true health condition of an older adult for long periods continuously without unduly inconveniencing the user. It is also a challenge to remain impervious to occasional external factors that can compromise the integrity of the data, such as motion artifacts or improperly worn sensors. This makes it even harder to validate the results from any new proposed analytical techniques and push the boundaries of longitudinal studies.

For information retrieval and mining of large-scale data, state-of-the-art database techniques should be employed to optimize the structure of data storage and accelerate the process of information retrieval. Cloud storage must also be taken into consideration for storing very-large-scale data, as long as the privacy and security issues of cloud storage can be addressed successfully. There are research opportunities here to develop feasible and scalable data organization and mining techniques. This also ties into data delivery, wherein the health-care provider must be able to quickly and easily access the required subset of user information from a large data set.

Information delivery paradigms and visualization

Given the human-centered nature of smart homes, information delivery to the user must remain seamless and effective. In smart home environments, the raw data collected via different sensors are overwhelming and may require domain-specific knowledge, which will introduce challenges in terms of data interpretability. Older adults with potentially diminished cognitive ability and scarce domain knowledge will be challenged to understand the overwhelming quantity of data. Moreover, the information delivery system may need to provide information not only to the care recipients but also to their caregivers, clinicians, and family members. This necessitates novel summarization techniques that leverage advanced algorithms to convert raw data into relevant, customizable, and comprehensible summaries for the different viewers. This provides the care recipients with insight into their health conditions and the caregivers the information to make knowledgeable clinical decisions.

An information delivery system typically consists of two components: algorithms for summarizing the information and an interface for information delivery. Data summarization is an important component, as the vast amounts of raw data from sensors need to be synthesized and formatted in such a way that the user can quickly and intuitively grasp actionable information. Summarization tools should provide sufficiently relevant information to the caregivers while considering the health conditions of the care recipients. These tools should not only work with large amounts of heterogeneous data and leverage machine-learning techniques but also remain cognizant of the clinical utility of the information delivered.

Furthermore, considering each care recipient's unique behavior can maximize the usefulness of the output; it is an essential task to design visualization tools that can deliver interpretable information in an accessible manner, especially for older adults with potentially diminished cognitive and sensory capacity. Thus, the paradigms must be thoughtfully designed with multiple pathways of delivery to robustly handle potential sensory impairments. Examples of different information delivery paradigms are shown in Figure 3.

When it comes to communicating the summarized information, the visualization formats can be quite diverse, ranging from a simple health score statistical visualization (e.g., plots and charts) to complicated 2-D or three-dimensional renderings of the complete smart home space. Thomas et al. present a suite of visualization tools called *PyViz* that uses algorithms to track the position of residents and provide an interactive graphical interface through which users can view the smart home system in real time and gain access to historical trends [28]. Chen et al. present a web-based visualization system (CASASviz) that takes this visualization technique one step further with a consumer-centric design [29]. Specifically, CASASviz applies data mining and machine-learning techniques to recognize user behavior patterns and detect unexpected changes that may be indicative of a decline in health status. Moreover, the visualization format of CASASviz can be customized to highlight the events of particular interest via a set of user-defined rules. Although CASASviz is among the earliest efforts to develop human- and consumer-centric visualization tools, research investigations on health visualizations from a consumer perspective remain scarce, especially for older adults. Age-dependent visualization has attracted research interest, taking normal, age-associated changes into consideration, such as deteriorated visual functions and reduced information processing efficiency [30]. For instance, graphical interfaces should remain as succinct as possible since older adults often have difficulties locating target items in a cluttered background. There is a research opportunity to explore intuitive delivery mechanisms beyond traditional displays. Visual information can be depicted in various forms, such as wall projections, smart lights, or even a single light-emitting diode customizable for various applications. Besides visual representation, information can also be delivered in other forms, including audio feedback and vibrotactile feedback.

Consumer-centric and disorder-specific visualization tools remain largely unexplored. Older adults with diverse health conditions and disorders may require more degrees of freedom to customize the information delivery according to their needs and capabilities. They may also want to prioritize viewing information that is relevant to their specific condition. However, the information delivery via current visualization techniques is generally fixed and ad hoc, and thus cannot be tailored to the specific requirements of different groups of older adults.

Finally, one important challenge is to cater to specific needs and display only the information of particular interest excluding redundant information at the right time. The information delivery methods must remain context-aware, which can help circumvent challenges associated with information overload that can ultimately lead to insensitivity to the information presented and negatively impact outcomes.

As previously mentioned, the privacy of personal information is a major concern in the context of smart homes for aging in place. Therefore, in any discussion of user interface and information delivery mechanisms we must also consider the privacy of the user. While there are many sophisticated algorithms and encryption mechanisms, care should be taken to ensure the right balance between protection of data and ease of use for senior citizens as well as any potential caregivers. User interfaces, like the one designed by Sivaraman et al. in which the user can choose between different security and privacy settings for different household devices, might be one of the solutions [22]. However, it can be argued that overzealous protection of information might hamper the overall goal of a health-monitoring smart system if users are not careful about the choices they make. So the challenge is not only to design user-friendly interfaces to protect privacy but also to provide proper privacy awareness among users.

An important research effort is the development of data obfuscation techniques to protect the fundamental privacy

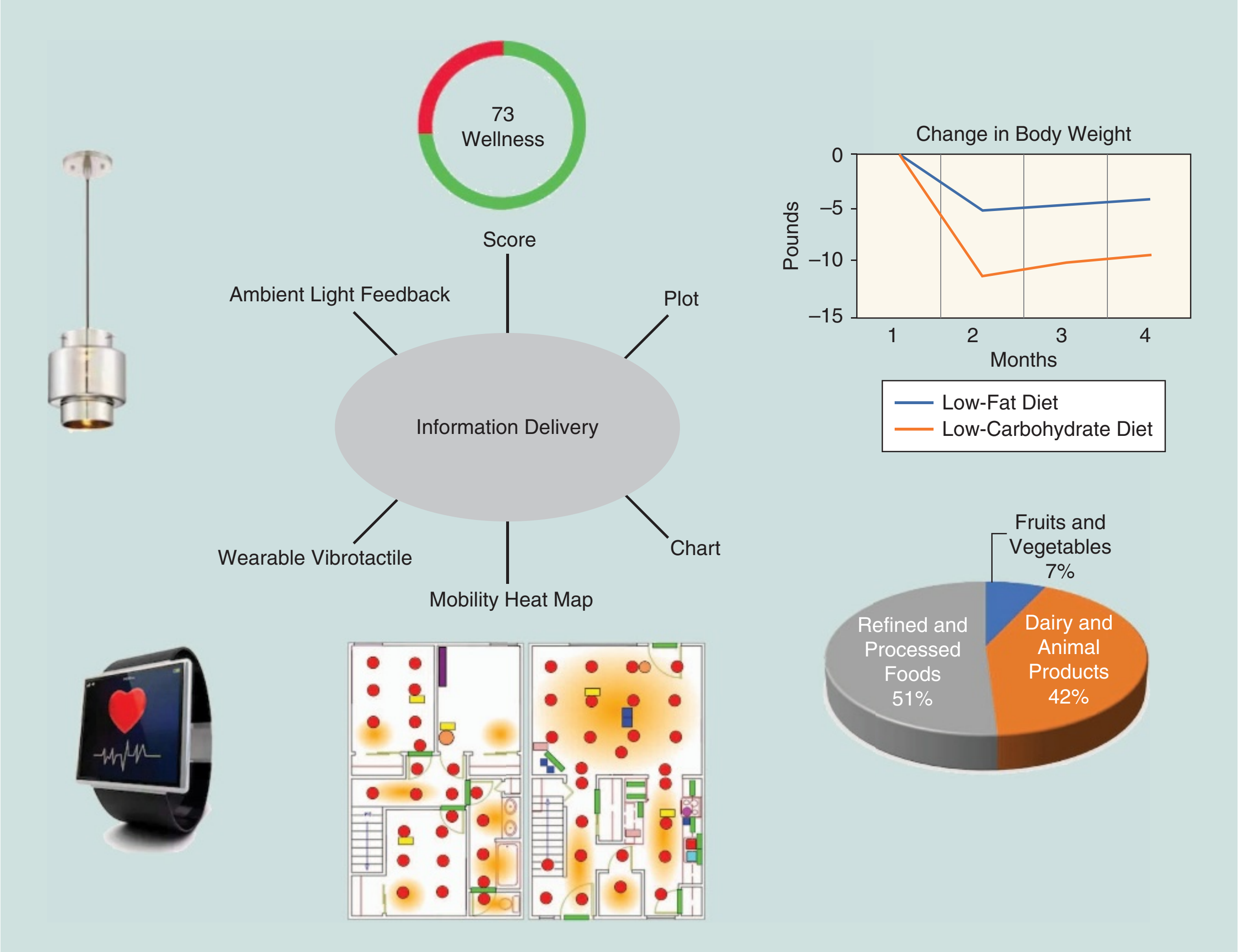


FIGURE 3. Different forms of information delivery. (The mobility heat map is from [29] and used with permission. Ambient light and wearable vibrotactile images courtesy of www.shutterstock.com.)

rights of the user while still providing health-care providers with sufficient actionable data.The use of lightweight authorization and encryption techniques for battery-operated devices is another research opportunity in this regard. Researchers should also consider designing contextual privacy-protection interfaces and devices to improve user discretion while keeping the balance between protection of personal information and the performance of the system in terms of achieving its goals. For example, the privacy-protection mechanism for an online purchase that could be viewed as typical should be different from that used for a transaction that would reveal important information about the user.

Conclusions

In this article, we identified some of the research challenges and opportunities associated with the key aspects of a smart home system with the objective of enabling aging in place. Many of the known problems in the fields of smart home sensing, signal processing, analytics, and visualization require solutions that are cognizant of the specific needs of the elderly. We highlighted several relevant recent works in the area to give the readers a perspective on the current status, while also presenting the necessary future directions of research to realize the vision of aging in place.

Acknowledgments

This work was supported in part by the National Science Foundation under grants CNS-1150079, CNS-1738293, ECCS-1509063, and EEC-1648451. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations.

Authors

Viswam Nathan (viswamnathan@tamu.edu) received his B.S. and M.S. degrees in computer engineering from the University of Texas at Dallas in 2012 and 2015, respectively. He has completed his Ph.D. degree thesis work in computer engineering at Texas A&M University, College Station. His research interests include the design and development of wearable and reconfigurable health-monitoring devices and associated signal processing techniques.

Sudip Paul (sudip.paul@tamu.edu) received his B.Sc. degree in electrical and electronic engineering from the Bangladesh University of Engineering and Technology, Dhaka, in 2015. He is currently pursuing his Ph.D. degree in biomedical engineering at Texas A&M University, College Station. He received the Best Paper Award at the 2017 IEEE International Conference on Imaging, Vision, and Pattern Recognition. His research interests include motion sensing, motion biomechanics, wearable health monitoring, and biomedical signal processing.

Temiloluwa Prioleau (to9@rice.edu) received her B.S. and Ph.D. degrees in electrical engineering from the University of Texas at Austin and Georgia Tech, Atlanta, in 2010 and 2016, respectively. She is currently a postdoctoral fellow at Rice University, Houston, Texas, and has accepted the position of assistant professor at Dartmouth College, Hanover, New Hampshire, for 2019. She is a recipient of awards including the National Science Foundation Graduate Research Fellowship, Achievement Rewards for College Scientists Fellowship, and Best Student Paper at the 2014 IEEE Engineering in Medicine and Biology Conference. Her research interests include human-centric sensing and health monitoring using mobile and wearable technology.

Li Niu (ln7@rice.edu) received his B.S. degree in computer science from the University of Science and Technology of China, Hefei, in 2011 and his Ph.D degree in computer science from Nanyang Technological University, Singapore, in 2017. He is currently a postdoctoral researcher at Rice University, Houston, Texas. He has published more than 20 papers presented at conferences and in journals. His current research interests include computer vision and machine learning.

Bobak J. Mortazavi (bobakm@tamu.edu) received his B.S. degree in electrical engineering and computer science and his B.A. degree in applied mathematics from the University of California, Berkeley, in 2007 and his Ph.D. degree in computer science from the University of California, Los Angeles, in 2014. He is currently an assistant professor of computer science and engineering at Texas A&M University, College Station. He previously conducted postdoctoral research at the Yale School of Medicine in the Yale Center for Outcomes Research and Evaluation, New Haven, Connecticut, on machine-learning for cardiovascular care and clinical outcomes. His research includes the intersection of computer science, electrical engineering, and medicine by integrating embedded systems, machine learning, and clinical outcomes research. He is a member of the Center for Remote Health Technologies and Systems.

Stephen A. Cambone (stevecambone@tamu.edu) received his undergraduate degree from the Catholic University of America, Washington, D.C., and his M.A. and Ph.D. degrees from Claremont Graduate School, California. He is currently the associate vice chancellor for Cybersecurity Initiatives, Texas A&M University System, College Station. He also holds the positions of director at the Institute for National Security and Cybersecurity Education and Research and professor of practice in the College of Engineering, at Texas A&M University. He is responsible for creating and leading an interdisciplinary cybersecurity program. In this role, he works with research professors and professionals in the schools and centers of the flagship campus, Texas A&M, and with the member universities of the Texas A&M University System.

Ashok Veeraraghavan (vashok@rice.edu) received his B.S. degree in electrical engineering from the Indian Institute of Technology, Madras, in 2002 and his M.S. and Ph.D. degrees from the Department of Electrical and Computer Engineering at the University of Maryland, College Park, in 2004 and 2008, respectively. He is currently an associate professor of electrical and computer engineering at Rice University, Houston, Texas. Prior to joining Rice University, he spent three years as a research scientist at Mitsubishi Electric Research Labs in Cambridge, Massachusetts. His research interests include computational imaging, computer vision, and robotics and their applications to personalized and mobile health care.

Ashutosh Sabharwal (ashu@rice.edu) received his B.Tech. degree from the Indian Institute of Technology Delhi in 1993 and his M.S. and Ph.D. degrees from The Ohio State University, Columbus, in 1995 and 1999, respectively. He is currently a professor in the Department of Electrical and Computer Engineering, Rice University, Houston, Texas. He received the 1998 Presidential Dissertation Fellowship Award, the 2017 Jack Neubauer Memorial Award, and the 2018 Advances in Communication Award. His research interests include information theory, communication algorithms, and experiment-driven design of wireless networks, as well as biobehavioral sensing to measure human behavior and its impact on human biology.

Roozbeh Jafari (rjafari@tamu.edu) received his B.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, and his Ph.D. degree in computer science from the University of California, Los Angeles, and completed a postdoctoral fellowship at the University of California, Berkeley. He is currently an associate professor of biomedical engineering, computer science and engineering, and electrical and computer engineering at Texas A&M University, College Station. He was the recipient of the National Science Foundation CAREER Award in 2012, the IEEE Real-Time and Embedded Technology and Applications Symposium Best Paper Award in 2011, and the Andrew P. Sage Best Transactions Paper Award from the IEEE Systems, Man and Cybernetics Society in 2014. He is an associate editor of *IEEE Sensors Journal*, *IEEE Internet of Things Journal*, and *IEEE Journal of Biomedical and Health Informatics*. His research interests include wearable computer design and signal processing.

References

[1] United Nations. (2015). World populating ageing 2015. New York. [Online]. Available: http://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2015_Report.pdf

[2] FQHC Germane. (2017). Baby boomers all grown up—The impact of the aging population on healthcare. [Online]. Available: https://www.fqhc.org/blog/2017/8/30/baby-boomers-all-grown-up-the-impact-of-the-aging-population-on-healthcare#_ftn4

[3] M. Amiribesheli, A. Benmansour, and H. Bouchachia, “A review of smart homes in healthcare,” *J.Ambient Intell. Humanized Comput.,* vol. 6, no. 4, pp. 495–517, 2015.

[4] L. Liu, E. Stroulia, I. Nikolaidis, A. Miguel-Cruz, and A. R. Rincon, “Smart homes and home health monitoring technologies for older adults: A systematic review,” *Int. J. Med. Informat.*, vol. 91, pp. 44–59, July 2016.

[5] N. K. Suryadevara, S. C. Mukhopadhyay, R. Wang, and R. K. Rayudu, “Forecasting the behavior of an elderly using wireless sensors data in a smart home,” *Eng. Appl. Artif. Intell.*, vol. 26, no. 10, pp. 2641–2652, Nov. 2013.

[6] H. Ghayvat, S. Mukhopadhyay, X. Gui, and N. Suryadevara, “WSN- and IoT-based smart homes and their extension to smart buildings,” *Sensors*, vol. 15, no. 5, pp. 10,350–10,379, 2015.

[7] A. Fleury, M. Vacher, and N. Noury, “SVM-based multimodal classification of activities of daily living in health smart homes: Sensors, algorithms, and first experimental results,” *IEEE Trans. Informat. Technol. Biomed.*, vol. 14, no. 2, pp. 274–283, Mar. 2010.

[8] S.-C. Kim, Y.-S. Jeong, and S.-O. Park, “RFID-based indoor location tracking to ensure the safety of the elderly in smart home environments,” *Personal Ubiquitous Comput.*, vol. 17, pp. 1699–1707, Dec. 2013.

[9] L. Wang, T. Gu, X. Tao, H. Chen, and J. Lu, “Recognizing multi-user activities using wearable sensors in a smart home,” *Pervasive Mobile Comput.*, vol. 7, pp. 287–298, June 2011.

[10] G. Mokhtari, Q. Zhang, G. Nourbakhsh, S. Ball, and M. Karunanithi, “BLUESOUND: A new resident identification sensor—Using ultrasound array and BLE technology for smart home platform,” *IEEE Sensors J.*, vol. 17, pp. 1503–1512, Mar. 2017.

[11] P. N. Dawadi, D. J. Cook, and M. Schmitter-Edgecombe, “Automated cognitive health assessment from smart home-based behavior data,” *IEEE J. Biomed. Health Informat.*, vol. 20, pp. 1188–1194, July 2016.

[12] S. Abbate, M. Avvenuti, P. Corsini, J. Light, and A. Vecchio, “Monitoring of human movements for fall detection and activities recognition,” in *Elderly Care Using Wireless Sensor Network: A Survey*, Y. K. Tan and G. Merrett, Eds. London: InTech, 2010, pp. 147–166.

[13] K. Williams and S. Kemper, “Exploring interventions to reduce cognitive decline in aging,” *J. Psychosocial Nursing Mental Health Serv.*, vol. 48, pp. 42–51, May 2010.

[14] H. G. Kang, D. F. Mahoney, H. Hoenig, V. A. Hirth, P. Bonato, I. Hajjar, L. A. Lipsitz, and N. Null, “In situ monitoring of health in older adults: Technologies and issues,” *J. Amer. Geriatrics Soc.*, vol. 58, no. 8, pp. 1579–1586, 2010.

[15] T. Hayes, M. Pavel, and J. Kaye, “An approach deriving continuous health assessment indicators from in-home sensor data,” *Assistive Technol. Res. Series*, vol. 21, pp. 130–137, 2008.

[16] A. D. Wood, J. A. Stankovic, G. Virone, L. Selavo, Z. He, Q. Cao, T. Doan, Y. Wu, L. Fang, and R. Stoleru, “Context-aware wireless sensor networks for assisted living and residential monitoring,” *IEEE Netw.*, vol. 22, pp. 26–33, July 2008.

[17] S. T. Peek, E. J. Wouters, J. van Hoof, K. G. Luijkx, H. R. Boeije, and H. J. Vrijhoef, “Factors influencing acceptance of technology for aging in place: A systematic review,” *Int. J. Med. Informat.*, vol. 83, no. 4, pp. 235– 248, 2014.

[18] F. D. Davis, “Perceived usefulness, perceived ease of use, and user acceptance of information technology,” *MIS Quart.*, vol. 13, pp. 319–340, Sept. 1989.

[19] J. Austin, H. H. Dodge, T. Riley, P. G. Jacobs, S. Thielke, and J. Kaye, “A smart-home system to unobtrusively and continuously assess loneliness in older adults,” *IEEE J. Transl. Eng. Health Med.*, vol. 4, pp. 1–11, June 2016.

[20] B. L. R. Stojkoska and K. V. Trivodaliev, “A review of Internet of Things for smart home: Challenges and solutions,” *J. Cleaner Prod.*, vol. 140, pp. 1454–1464, Jan. 2017.

[21] S. L. Keoh, S. S. Kumar, and H. Tschofenig, “Securing the Internet of Things: A standardization perspective,” *IEEE Internet Things J.*, vol. 1, pp. 265–275, June 2014.

[22] V. Sivaraman, H. H. Gharakheili, A. Vishwanath, R. Boreli, and O. Mehani, “Network-level security and privacy control for smart-home IoT devices,” in *Proc. 11th IEEE Int. Conf. Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2015, pp. 163–167.

[23] D. He and S. Zeadally, “An analysis of RFID authentication schemes for Internet of Things in healthcare environment using elliptic curve cryptography,” *IEEE Internet Things J.*, vol. 2, pp. 72–83, Feb. 2015.

[24] H. Zheng, H. Wang, and N. Black, “Human activity detection in smart home environment with self-adaptive neural networks,” in *Proc. IEEE Int. Conf. Networking, Sensing and Control*, 2008, pp. 1505–1510.

[25] A. M. Tabar, A. Keshavarz, and H. Aghajan, “Smart home care network using sensor fusion and distributed vision-based reasoning,” in *Proc. 4th ACM Int. Workshop Video Surveillance and Sensor Networks (VSSN)*, 2006, pp. 145–154.

[26] P. Dawadi, D. J. Cook, and M. Schmitter-Edgecombe, “Smart home-based longitudinal functional assessment,” in *Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp)*, 2014, pp. 1217–1224.

[27] S. Hagler, D. Austin, T. L. Hayes, J. Kaye, and M. Pavel, “Unobtrusive and ubiquitous in-home monitoring: A methodology for continuous assessment of gait velocity in elders,” *IEEE Trans. Biomed. Eng.*, vol. 57, pp. 813–820, Apr. 2010.

[28] B. L. Thomas and A. S. Crandall, “A demonstration of PyViz, a flexible smart home visualization tool,” in *Proc. IEEE Int. Conf. Pervasive Computing and Communications Workshops (PERCOM Workshops)*, Mar. 2011, pp. 304–306.

[29] C. Chen and P. Dawadi, “CASASviz: Web-based visualization of behavior patterns in smart environments,” in *Proc. IEEE Int. Conf. Pervasive Computing and Communications Workshops (PERCOM Workshops)*, Mar. 2011, pp. 301–303.

[30] U. Backonja, N.-C. Chi, Y. Choi, A. Hall, T. Le, Y. Kang, and G. Demiris, “Visualization approaches to support healthy aging: A systematic review,” *J. Innovation Health Informat.*, vol. 23, no. 3, pp. 600–610, 2016.

SP

Yuan He, Junchen Guo, and Xiaolong Zheng

From Surveillance to Digital Twin

Challenges and recent advances of signal processing for the industrial Internet of Things

INTERNET OF THINGS—ISTOCKPHOTO.COM/IAREMENKO
CIRCUITS—IMAGE LICENSED BY INGRAM PUBLISHING

With the recent advances in the Internet of Things (IoT), the significance of information technologies to modern industry is upgraded from purely providing surveillance-centric functions to building a comprehensive information framework of the industrial processes. Innovative techniques and concepts emerge under such circumstances, e.g., digital twin, which essentially involves data acquisition, human–machine-product interconnection, knowledge discovery and generation, and intelligent control, etc. Signal processing techniques are crucial to the aforementioned procedures but face unprecedented challenges when they are applied in the complex industrial environments. In this article, we survey the promising industrial applications of IoT technologies and discuss the challenges and recent advances in this area. We also share our early experience with Pavatar, a real-world industrial IoT system that enables comprehensive surveillance and remote diagnosis for ultrahigh-voltage converter station (UHVCS). Potential research challenges in building such a system are also categorized and discussed to illuminate the future directions.

Introduction

With the prosperity of various embedded sensors, low-power wireless communications, and efficient signal processing techniques, the IoT has achieved explosive development and proliferation in recent years. The IoT offers opportunities to bridge the physical world and cyberspace, enabling fine-grained sensing of objects and environments, continuous data gathering, comprehensive information fusion, deep analysis, and real-time feedback or control over the connected targets. According to Gartner's report, there are approximately 8 billion connected things providing smart services in our daily life, for example, in assisted living, building surveillance, traffic control, and environment monitoring, etc [1].

The ever-developing IoT attracts the interest of both industry and academia. Technology giants have already taken their steps. Huawei, one of the world's largest telecom equipment makers, has devoted itself to standardizing the narrow-band IoT (NB-IoT) as the next generation of low-power wide-area networks to

Digital Object Identifier 10.1109/MSP.2018.2842228
Date of publication: 28 August 2018

fulfill its long-term strategy for building a better-connected world [2]. Practical IoT platforms have also been vigorously promoted recently, e.g., Android Things (Google), Predix [General Electric (GE)], Azure IoT Suite (Microsoft), etc. In the meantime, academia focuses on exploring cutting-edge techniques to boost the application and development of the IoT, such as wireless sensing, indoor localization, low-power networking, backscatter communication, visible light communication, mobile computing, edge computing, privacy and security, etc.

Among all of the promising scenarios, applying IoT technologies in modern industry has great potential and practical significance. In 2011, Industry 4.0 is proposed to equip traditional manufacturing with cyberphysical systems to start a new industrial revolution. GE formally put forward the concept of the Industrial Internet in 2012 [3]. GE then established the Industrial Internet Consortium with AT&T, Cisco, Intel, and IBM, bringing together the world's leaders in the manufacturing, telecom, networking, semiconductor, and computer industries, respectively, to promote the industrial IoT systems.

Due to the prosperity of IoT techniques in the past few years, digital twin has recently gained extensive attention. Digital twin represents a dynamic digital replica of physical assets, processes, and systems, which comprehensively monitors their whole life cycle. The backbone technology of digital twin is the IoT for real-time and multisource data collection. In addition, it integrates artificial intelligence and software analytics to create digital simulation models that dynamically update and change along with their physical counterparts. Moreover, digital twin adopts modern data visualization schemes such as virtual reality (VR) and augmented reality that can provide more illustrative and user-friendly views.

Therefore, compared to traditional surveillance systems, digital twin provides more sensing modalities with stricter timeliness guarantees, and integrates more intelligent data analysis and friendlier display and interaction. With digital twin, we can not only better understand and predict the performance of machines and systems, but also optimize business operations for equipment suppliers and consumers. However, it is a nontrivial task to achieve such comprehensive monitoring along with requirements such as timeliness, accuracy, scalability, and interoperability in industrial IoT. We summarize potential research challenges as follows.

- First, digital twin pushes the boundary of sensing capabilities toward the physical world. Sensing methods that monitor diverse physical metrics but rely on less resources are deemed to be more practical in industrial IoT. Wireless and battery-free sensing integrating efficient techniques of data cleaning and signal processing can support lightweight and robust monitoring. How to extend the sensing capabilities of wireless signals [4]–[7] and how to refine the sensing precision from vulnerable readings [8]–[10] have triggered numerous research motivations over the past few years.
- Second, visual sensing is extremely informative for the surveillance of physical assets and their surroundings. In digital twin, intensive networked cameras are deployed at a high density to provide seamless monitoring. On one hand, processing intensive networked videos need the upgrade of computing architecture for timeliness requirements, e.g., collaborative edge computing [11]–[13]. On the other hand, enabling a resource-constrained IoT device with modern analysis techniques, e.g., deep learning, can also release the pressure of cloud infrastructure and save the network bandwidth [14]–[17].
- Third, new forms of communication and networking is anticipated in digital twin for efficient data transmission. Recent advances in low-power wireless networking such as low-power wide-area networks [2], parallel backscatter transmissions [18], and software-defined low-power wireless [19] has drawn much attention. In this section, we emphasize the research challenges and opportunities on direct communication among heterogeneous wireless technologies that share the same frequency band [20]–[23], and their upper-layer protocols as well as applications [24]–[26].
- Last but not least, comprehensive data analysis and system diagnosis need innovative and dedicated signal processing methods. For example, anomaly detection and repairing of time-series data [27], feature selection from heterogeneous stream data [28], and fault analysis based on incomplete data [29] should also be well designed.

Practical industrial IoT and signal processing

Signal processing algorithms are indispensable in almost every layer of industrial IoT. In this section, we survey the most recent research works and corresponding signal processing techniques, to provide an overview of the current progress from sensing, networking to data analysis in industrial IoT.

Wireless and battery-free sensing

In practical industrial scenarios, many physical metrics need to be closely monitored, such as temperature and humidity, vibration and noise, rotation speed, liquid leakage, etc. Although the advances of modern sensor technologies enable the sensibility of more metrics, a part of these metrics cannot be provided due to the complicated operational environments in real-world deployments that have the special characteristics that are given next.

- *Requirement of nonintrusive sensing*: Adding dedicated sensors into the existing equipments costs too much because these intrusive sensors may trigger hardware updates or even redesigns. Hence, nonintrusive sensing methods are more preferred.
- *Large-scale sensing targets*: The large number of targets to be monitored makes it unaffordable to deploy dedicated sensors at all the monitoring points. Novel low-cost sensing solutions are desired.
- *Limited sensing capability*: Physical metrics can be very fast changing, but most nonintrusive sensors can usually provide undersampled data. How to fill this gap remains a challenge.

Because traditional sensors are mostly intrusive, those approaches cannot be deployed with an operational machine that hasn't been initially equipped with such a capability. Other high-resolution approaches, e.g., cameras and lasers, suffer from the line-of-sight problem and are restricted in the application context.

Moreover, audio-based sensing is sensitive to environmental noises, which is therefore impractical for real-world industrial applications. Wireless and battery-free sensing, e.g., radio-frequency identification (RFID), which leverages backscattered radio-frequency signals to carry information, has received plenty of attention in the past few years, due to its low-cost, nonintrusive, and easy-deployment properties. A typical RFID system, as shown in Figure 1, consists of RFID tags that store information in nonvolatile memories, and two-way radio transmitter–receivers called *RFID readers* that send signals to tags and receive their responses.

Recent advances in RFID offer a promising technique for cross-modal sensing where many physical metrics are sampled with only battery-free devices and wireless signals [4]–[7]. In the meantime, the resolution of RFID sensing—especially battery-free localization and tracking—has been well improved over the past few years [8]–[10].

Cross-modal sensing with RFID

Apart from parsing the information encoded in backscatter signals from tags, widely employed commercial RFID readers, e.g., ImpinJ Speedway R420, Alien ALR-9900, and Zebra FX9500, can interrogate the readings of received signal strength indicator (RSSI) and phase values at the frequency of approximately 40 Hz. The changes in RSSI and phase offer space for the cross-modal sensing of other physical metrics, e.g., vibration [4] and eccentricity [6] of rotating machines, liquid category [5], and human–object interaction [7]. However, the relatively low interrogating frequency offered by commercial readers brings in additional research challenges in industrial scenarios.

A recent battery-free work, TagBeat, offers inexpensive and pervasive cross-modal sensing of mechanical vibration frequency with commercial RFID devices [4]. The phase shifts caused by micro vibration are too tiny to distinguish, and the high-frequency vibration is hard to capture with the limited-frequency readings. Thus, TagBeat first magnifies weak vibration signals without losing their features and then leverages compressive sensing (CS) to recover the high-frequency signals with the low-frequency samplings. To guarantee safety, another work, TagScan [5], utilizes the differences of RF signals when traversing different kinds of liquid to classify them. In this work, a feature that only relates to the liquid material is extracted from RSSI and phase values with a signal propagation and attenuation model.

High-precision RFID localization and tracking

In industrial automation, object localization and tracking is one of the most critical demands. Wireless and battery-free backscattering offers a lightweight and low-cost solution for localization and tracking of the materials in warehouses and products on production lines. Early works achieving a median accuracy of tens of centimeters either rely on RSSI for distance estimation and fingerprint map construction, or calculate the angle of arrival (AoA) for continuous localization. Recent proposals integrate reference tags or antenna arrays to calculate phase changes for centimeter-scale precision. Here we survey the most recent works on RFID localization and tracking, which improve not only the task precision but also the robustness and the practicability of sensing systems [8]–[10].

A recent work, OmniTrack [8], solves the problem of the precision degradation caused by the phenomenon of the antenna polarization when the orientation of a RFID tag changes. To achieve centimeter-level localization and orientation of a mobile tag, OmniTrack models the linear relationship between the tag orientation and the phase change of the backscattered signal. To deal with high-noise and complicated multipath environments and to soften the deployment restricts of antennas, Xiao et al. propose a double-tag system for accurate and robust object localization and tracking [9]. The work demonstrates that the phase difference of closely deployed double tags can effectively exclude the impact of undesired signals such as device noises and multipath interferences. RFind [10] manages to use time-of-flight (ToF) for RFID tag localization. To

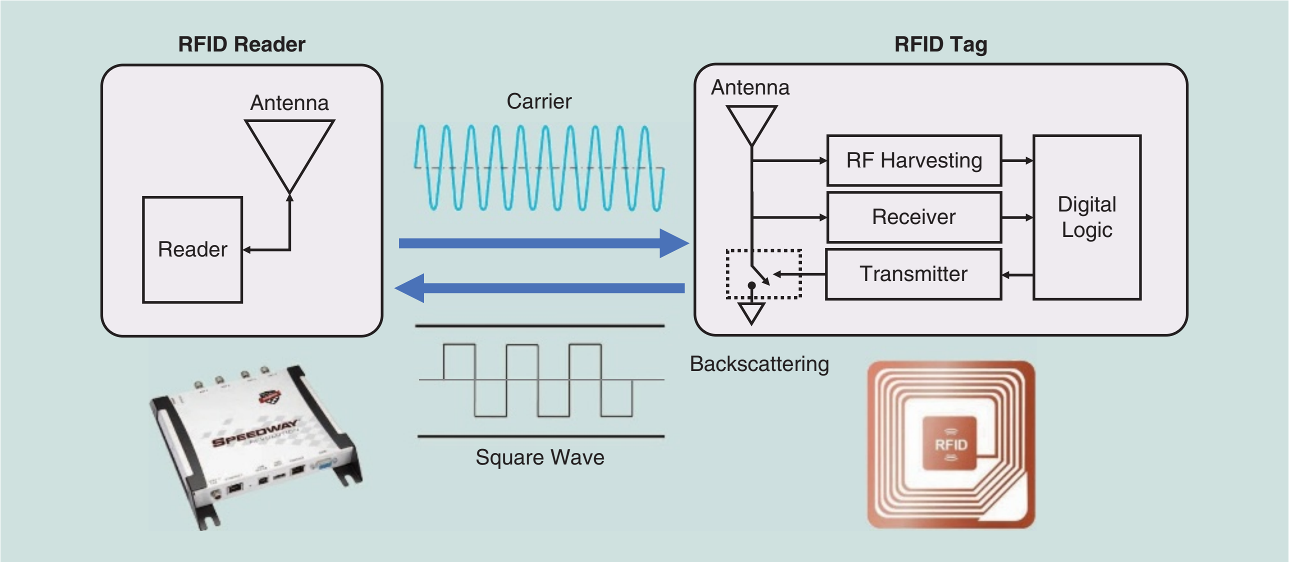


FIGURE 1. Backscattering with RFID systems.

achieve subcentimeter accuracy, a very large bandwidth of multiple gigahertz is often needed in ToF-based methods, which, however, is not compliant with Federal Communications Commission (FCC) regulation and RFID protocols. Thus, RFind generates a virtual ultrawide bandwidth by importing extremely low-power but efficient hopping localization frequencies outside the industrial scientific medical (ISM) band while keeping the normal communication band for powering up commercial tags.

To summarize, battery-free RFID sensing offers a new paradigm that not only can measures specific physical metrics with just wireless signals, but also can provide high-precision results. Except extending physical modalities, improving resolution, timeliness, and reliability of battery-free RFID sensing offers prime candidates for further studies. Besides, new nonintrusive wireless tag systems are increasingly gaining more attention recently, e.g., LiveTag [7] designs multiple metallic structures of a Wi-Fi tag to disturb ambient Wi-Fi channels for information expression. Further, it leverages customized multiantenna beamforming algorithms to sense the human–object interaction. Moreover, we will show our preliminary explorations of designing RFID systems for real-world industrial IoT in the section "Case Study: Pavatar."

Visual sensing from intensive networked videos

Surveillance cameras are one of the most commonly used IoT devices in industrial IoT because the visual sensing provides numerous informative clues. In modern industries, cameras are deployed with a high density to seamlessly monitor the status of machines and the activities of workers. The characteristics of visual sensing in industrial IoT is as follows.

- *Timeliness requirement*: Video analysis usually has a strict requirement of timeliness in modern industries. How to fulfill the real-time processing on resource-limited devices while reducing the transmission latency remains a challenge.
- *Information sparsity*: Camera surveillance systems generate intensive video data, but the spatiotemporal sparsity of significant information needs efficient processing.
- *Seamless cooperation*: Visual clues provided by one single camera is partial and limited, thus seamless cooperation among the networked cameras is desired to perform complicated sensing tasks.

Visual sensing applications on a large-scale camera network need not only the optimized allocation of the computation resources but also the efficiency and the accuracy of the vision tasks. In this section, we first introduce a rising computation paradigm, edge computing for multimedia IoT data processing [11]–[13] and then discuss efficient and accurate video analysis algorithms of resource-constrained embedded devices [14]–[17].

Edge computing for large-scale networked video processing

The networked cameras are expected to cooperate for a comprehensive understanding of the monitoring targets. However, uploading all of the multimedia data stream to the cloud is infeasible due to its limited processing capacity of the cloud, the unpredictable latency induced by the network transmission, and the unaffordable cost of the network bandwidth. Edge computing, a new computation paradigm between embedded computing and cloud computing, performs data processing and analyzing at the edge of networks. Large-scale networked video analytics is considered the killer app of edge computing [11].

Recent practical video analytics systems start adopting edge computing to deal with large-scale networked video, although there has not yet been a universally standard architecture. In [11], a practical system rocket for traffic monitoring in Bellevue, Washington, is proposed to discuss potential prospects of edge computing for the surveillance video processing. Model predictive control is used to allocate limited computation and network resources between the edge servers and the cloud server. A recent edge-computing architecture [12] introduces another offloading mode, where multiple edge servers cooperatively serve one camera and build a performance model with the compression ratio as the input. Then it separates the NP-hard problems of the edge server selection and the compression ratio selection, and solves them with heuristic algorithms. Besides offloading and scheduling, information sparsity can be leveraged to reduce computation costs among resource-constrained edge servers. The recent work ViTrack [13] proposes a spatiotemporal CS algorithm to recover the camera-level trajectories for the monitored vehicles by processing just 1/50 of the raw frames.

Practical video analytics with embedded deep learning

Recent advances in deep learning, especially convolutional neural networks (CNNs), have pushed the boundaries of computer vision. Basically, existing CNN applications purely rely on cloud infrastructures. However, problems such as network transmission delay, expensive but limited bandwidth, user privacy and costs

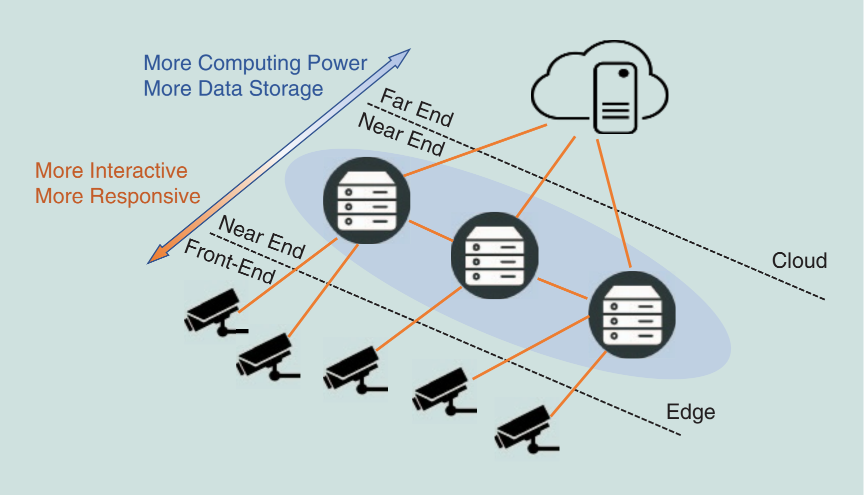


FIGURE 2. Edge computing for large-scale networked video processing.

of high-performance cloud servers make cloud-based solutions infeasible for large-scale video analysis in industrial scenarios. One potential trend to solve this dilemma is to enable real-time deep learning directly on end devices.

In a typical CNN model, convolutional layers that extract features consume much of the executing time because of the window-by-window convolutional operations, while fully connected layers that conduct the classification tasks take up much of the model weights because of the dense connections among neurons. Thus, to satisfy the requirement of the low-latency performance, we can adopt different strategies to optimize different modules, e.g., the structure pruning for the deep models [14], [15] and the runtime optimization of the inference frameworks [16], [17]. Model structure pruning methods such as DyNS [14] and DeepIoT [15] try to eliminate the redundancy in the model parameters through a three-step procedure: importance estimation, parameter pruning, and model retraining. Unimportant parameters are pruned to speed up computation and save storage space. Runtime

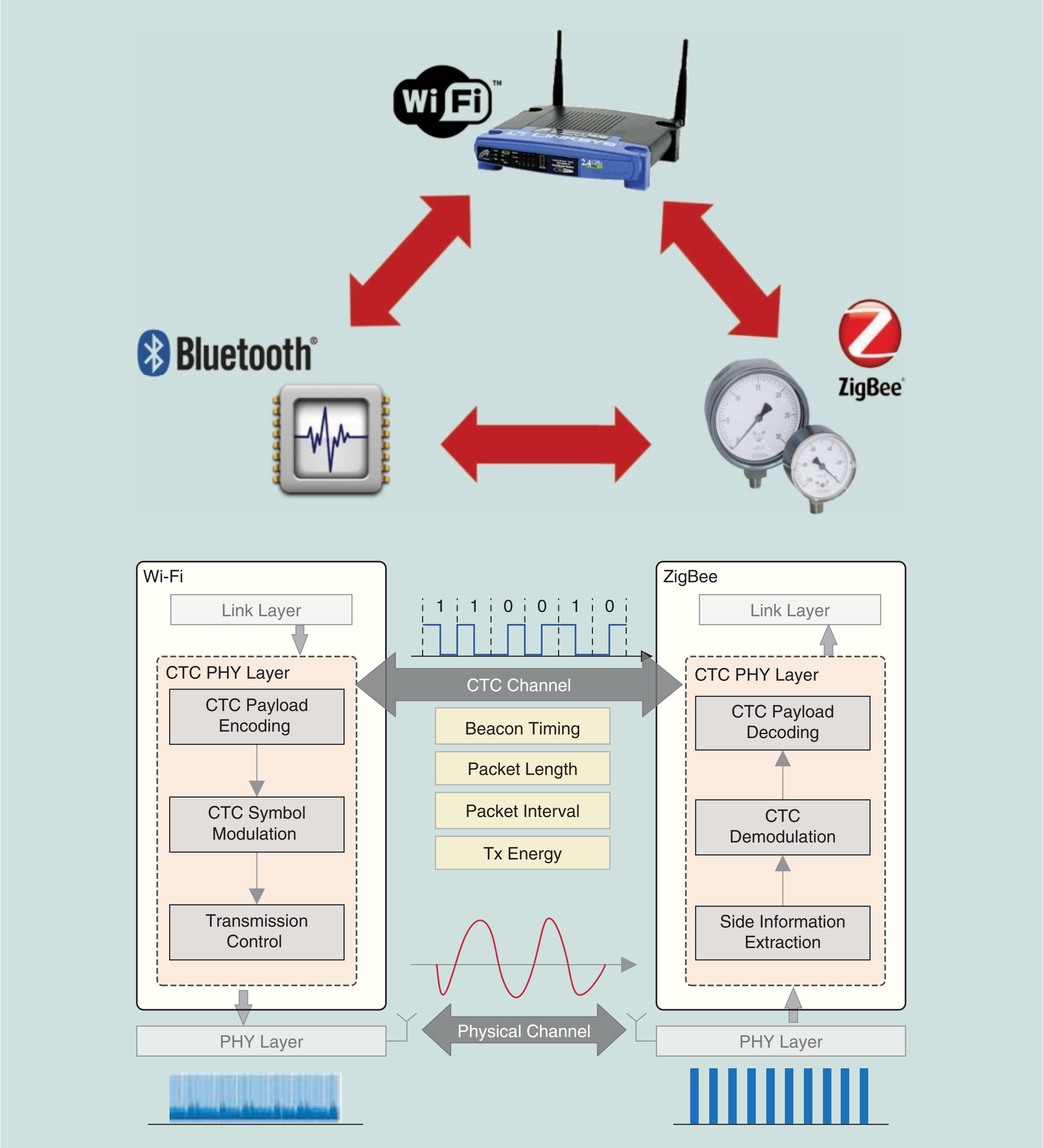


FIGURE 3. The architecture of CTC.

optimization utilizes computation parallelization and pipeline scheduling [16], intermediate-result caching [17], etc. to optimize the computation procedures of deep-learning inference frameworks on IoT devices.

In summary, edge computing is deemed as a promising architecture for practical visual sensing on ubiquitous surveillance cameras, and deep-learning algorithms with amazing analysis capabilities can be tailored to edge devices. In industrial IoT, effectiveness and timeliness are two dispensable, but mutually exclusive, performance indicators. Therefore, to maximize both of them, we believe enabling deep learning at the edge of networks is a promising direction.

Cross-technology heterogeneous wireless communication

In digital twin for smart factories, embedded sensors with various sensing capabilities are networked together to monitor the same area. These sensors might adopt heterogeneous wireless communication technologies, such as Wi-Fi, Bluetooth, ZigBee, and long-term evolution (LTE). The characteristics of heterogeneous networking environment are as follows.

- *Heterogeneous interference*: The majority of popular wireless technologies share the same frequency band, e.g., 2.4-GHz ISM band. Therefore, heterogeneous interferences and collisions are very likely to occur.
- *High-density deployment*: In many cases, networked sensors are densely deployed, which induces nontrivial challenges in collecting data in real time.
- *Interconnecting heterogeneous devices*: Due to the complicated operating states of industrial machinery, multiple devices need to exchange information in suit for a real-time understanding of current states.

Today, how to organize, manage, and cooperate heterogeneous IoT devices is increasingly drawing attention. A simple solution is to deploy a gateway with various radio interfaces for access control and information exchange among heterogeneous devices. Possible communication bottleneck and extra hardware cost drive researchers to explore the direct communication ability among different technologies, thus cross-technology communication (CTC) is proposed. With CTC, heterogeneous devices can directly exchange information for fast and effective control and cooperation, which perfectly satisfies the timeliness and interconnection requirements in industrial IoT.

The basic idea of CTC is that, although heterogeneous wireless technologies can't directly decode the packets from another technology, side-channel information of wireless transmissions, e.g., transmission time, beacon shifting [20], RSSI amplitude [21] etc., can be leveraged to encode bits. These corresponding methods are called *packet-level modulation* because one or more original packets should be transmitted to modulate one bit. Although recent CTC works manage to improve the throughput by deeply exploiting the coexistence environment to encode more bits simultaneously, packet-level modulation still offers a relatively low throughput, compared to the original wireless technologies.

Therefore, a new trend of CTC called *physical-level emulation* tries to emulate the heterogeneous signals directly in the physical layer to achieve a throughput comparable to the original wireless technology, e.g., up to 250 kilobits/s for ZigBee. WEBee [22] is the most representative work that meticulously fills the payload of a transmitted Wi-Fi frame to directly emulate ZigBee frames. The feasibility of the bit emulation is ensured by a redundancy coding technology of ZigBee called *direct sequence spread spectrum* (*DSSS*). The inevitable differences of the emulated chip sequences and the predefined chip sequences can be tolerated by the DSSS symbol matching.

The aforementioned CTC works have successfully established direct communication among heterogeneous wireless technologies in the physical layer. Protocols and applications can be further built upon these infrastructures. A basic scenario is to use CTC packets as a medium access control protocol for the channel coordination in coexisting environments, which is one of the primary intuitions of this emerging technique. ECC [24] introduces a cross-technology clear-to-send signal to negotiate an aggregated white space for better ZigBee communication. Moreover, Crocs [25] leverages CTC to directly synchronize Wi-Fi and ZigBee devices. To achieve a more robust and accurate time synchronization, Crocs first incorporates a short CTC beacon based on Barker code for a more accurate time alignment and then sends tim stamps via CTC transmissions. StripComm [26] applies CTC to a more densely coexisting environment and faces severe challenges with dynamic wireless interferences. To make the energy-encoding CTC more robust against unavoidable wireless interferences, StripComm encodes bits with Manchester coding, and decodes bits after the interference cancellation based on specific signal similarities.

In a nutshell, recent advances in CTC have experienced two stages, from packet-level modulation to physical-layer emulation. The validity and practicability of these approaches have been verified by the throughput comparable to the original wireless in [22]. Enhancing more modern wireless technologies, e.g., LTE and NB-IoT with the ability of CTC, as one of the future directions of CTC, faces the new challenges of the bandwidth asymmetry and the mismatch of the transmission rates. Moreover, facilitating the upper-layer standards and protocols to build cross-technology networks is also a very fascinating topic.

Data analytics

The data analytics layer plays a vital role in industrial IoT to provide smart services. The sensing layer samples raw data of physical metrics, the networking layer conveys data and, finally, the data analytics layer identifies patterns or mines the principles behind. The data analytics in industrial IoT have the following characteristics.

- *Low quality of raw data*: Due to the hardware imperfection or the unreliable wireless transmissions, the raw data generated by IoT devices are usually of low quality, which brings challenges for the accurate analytics.
- *Multisource data*: The data from multiple sensors may be redundant and even contradictory. Obtaining the truth from multisource data desires more advanced signal processing methods.

- *Partially labeled data*: In industrial scenarios, the high-frequency and continuous stream data is very difficult and impractical for manual labeling. Dealing with partially labeled data is also very challenging.

Analyzing IoT stream data with these characteristics is deeply associated with advanced signal processing algorithms, including data cleaning [27], feature selection [28], and event classification and system diagnosis [29].

Anomaly correction of time series data

Anomaly detection (or further anomaly correction of time series data) is an indispensable preprocessing step for upper-layer applications, such as event detection and fault diagnosis. In [27], Zhang, et al. suggest that simply filtering out anomalies will damage the continuity of time-series data, and the intermittent and incomplete time series would possibly affect subsequent classifiers. Different from existing rule-based repairing, e.g., the speed-constraint model and the autoregression model, an iterative minimum repairing (IMR) algorithm based on sparse-labeled ground truth is proposed. The sparse-labeled truth points, which can be obtained by a reliable sensor with a relatively long sampling period or manual labeling, can better fix continuous errors. Rather than sequentially repairing one error point for just one time, an IMR algorithm iteratively adjusts error points until the global convergence.

Data-driven feature selection

The multisource data can be redundant for upper-layer applications. Apart from the guidance of the physical models, data-driven feature selection can improve the final performance. In [28], Li et al. point out that traditional feature selection methods either consider only the informativeness of features regardless of sample labels, or are optimized for some particular classifiers. Hence, they leverage the sample labels and propose a novel information greedy feature filter (IGFF) method that is independent from the classifiers. With rigorous mathematical proofs, IGFF selects the optimal subset of features by maximizing mutual information between the candidate variables and the fault labels. The experiments on the real-world data set about air-handling units of a smart building shows that, regardless of back-end classifiers, IGFF can achieve a much higher improvement in the classification accuracy than the traditional methods and the empirical selection.

Event classification with partial labeled data

Fault detection is an event classification problem that classifies a short time series data from multiple sources into normality or particular faults. Current methods are mainly based on supervised learning. In industrial scenarios, however, the high-frequency and continuous stream data are almost unlabeled. Manual labeling by domain experts means considerable labor costs, which is impractical for real-world systems. In [29], a hidden structure semisupervised machine (HS^{3}M) is proposed to deal with sparsely labeled industrial IoT data. HS^{3}M incorporates fully labeled data, partially labeled data, and unlabeled data with a unified-format loss function, thus it can fully utilize all available data sets to learn a more generic model. Tested on an industrial IoT data set of a practical power distribution system, HS^{3}M can achieve at least 9% gain of accuracy and 10% gain of false positive in comparison to the runner-up method.

In summary, advanced signal processing technologies are indispensable to deal with fallible, multisource, and partially labeled industrial IoT data. Moreover, we believe practical data analytics is deeply associated with the characteristics of the target systems, which will be addressed in the next section.

Case study: Pavatar

In this section, we introduce our early experience with a real-world industrial IoT system, Pavatar [30]. Pavatar is an IoT system for UHVCS management. The UHVCS, built at the hub points of the ultrahigh-voltage power grid, efficiently performs dc/ac transformation of clean energy, e.g., wind, solar, water, and nuclear power. Globally connected UHVCSs are expected to construct the backbone of the Global Energy Internet (GEI), which is deemed to alleviate energy problems such as the exhaustion of fossil fuel, environmental pollution, and supply–demand imbalance. A large rotating machine called a *synchronous compensator* is the core component of an UHVCS. Its critical function is to stabilize the outgoing current by generating or absorbing reactive power, in response to unpredictable voltage fluctuations, and thus ensuring GEI's stability, safety, and reliability. Clearly, proper operation of synchronous compensators is of vital importance to GEI. There have been various conventional solutions for power plant monitoring, e.g., manual checking and video surveillance. However, those solutions are generally inefficient, inaccurate, and costly.

Our team collaborates with the State Grid Corporation of China to launch the Pavatar project in one UHVCS located in Hunan, China. Aiming to build a digital twin of this UHVCS, Pavatar monitors the entire operation process in real time and provides decisions and support for UHVCS administrators. The functionality of Pavatar generally includes the following key aspects:

- Comprehensive sensing of synchronous compensators and their cooling systems, operation environments, and surrounding human activities.
- Heterogeneous data visualization in the form of VR.
- System error prediction, anomaly detection, and root-cause diagnosis.

Figure 4 shows the architecture of Pavatar. Pavatar collects data from both built-in and ambient sensors in UHVCSs. Typical internal sensor readings include temperature, pressure, vibration, rotation, etc., which provide the key metrics for decision making. In the surrounding environment, low-power and battery-free sensors are deployed to sense temperature, humidity, noise, air quality, and liquid leakage, etc., as supplementary information. In addition, networked cameras are deployed to cover walkable areas. The maximum density of sensor deployment is about 50/m^2, the highest sampling frequency of internal sensors is around 10 KHz, and the total data volume size per day is over 1 TB. The high-frequency and big-volume stream data are collected and transmitted through heterogeneous networks to

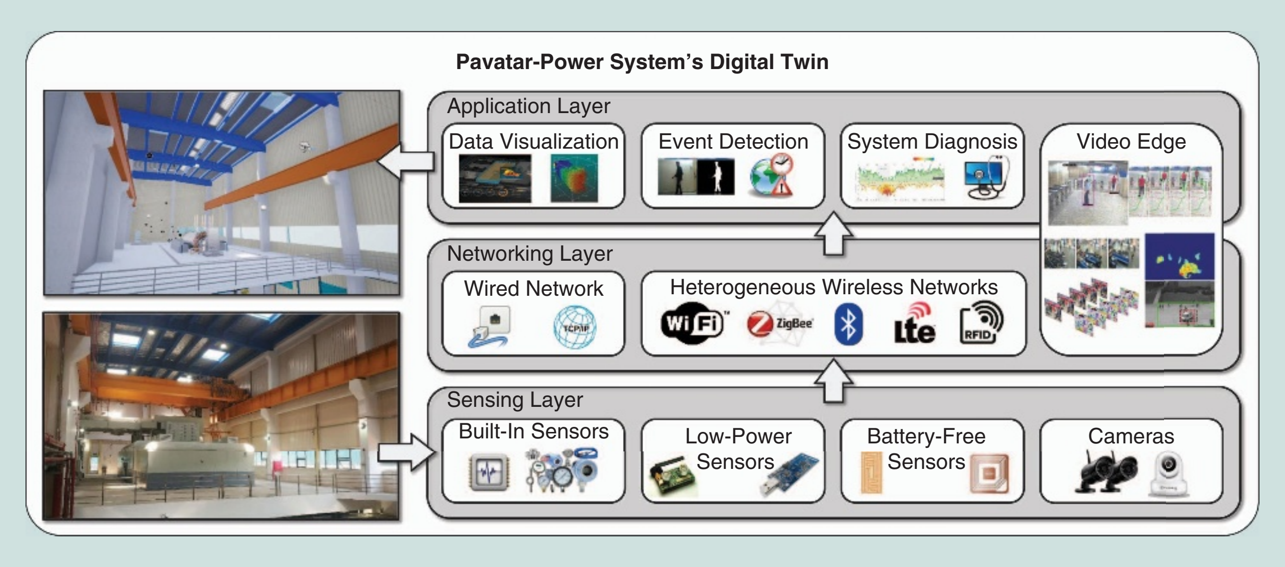


FIGURE 4. The architecture of Pavatar.

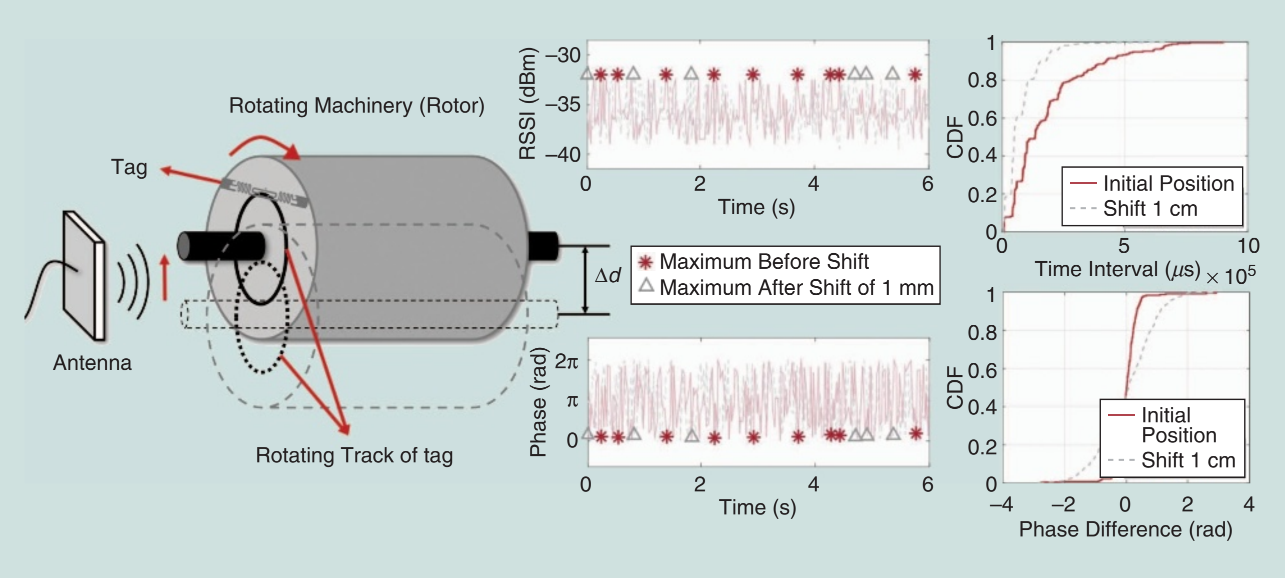


FIGURE 5. Due to the mismatch between the interrogation frequency and the rotation frequency, RED extracts statistic features for eccentricity (centroid shift) detection.

fulfill upper-level applications such as data visualization, event detection, and system diagnosis. Moreover, a three-layer edge-computing architecture is proposed to process massive video data. In the following, we present some of our recent works regarding Pavatar, which leverage advanced signal processing methods to deal with the problems of industrial IoT.

Battery-free sensing for eccentricity detection

Eccentricity, which stands for the displacement of rotating center, is essential for rotating machines, e.g., synchronous compensators in Pavatar. Traditional techniques based on special embedded sensors are either hard to deploy or not practical. Our recent work, RFID-based eccentricity detection (RED), proposes a battery-free RFID sensing system tailored to the classification of the eccentricity status [6]. As shown in Figure 5, RED first extracts features of statistic characteristics e.g., the cumulative distribution functions of the phase difference and the time interval between measured signal peaks, then constructs a Markov model to process stream data without training for a specific environment.

Parallel backscatter transmissions

RFID tags are deployed in Pavatar with the density up to 40/m^2 for liquid leakage detection. The dense deployments in industrial IoT require new networking techniques for efficient data collection. Thus, we recently proposed a practical system called *FlipTracer* that decodes collided signals to achieve reliable parallel backscatter transmissions [18]. We found that the transition of tag states is usually caused by the discrete signal flip of a single tag. Thus, instead of the direct classification, the states can be inferred by modeling the transition probabilities. As shown in Figure 6, FlipTracer constructs a one-flip graph (OFG) in the in-phase and quadrature (IQ) domain to model the transition patterns and then tracks the OFG to resolve the collided signals. FlipTracer is able to achieve an aggregated throughput of 2 megabits/s, which is six times higher than the existing methods.

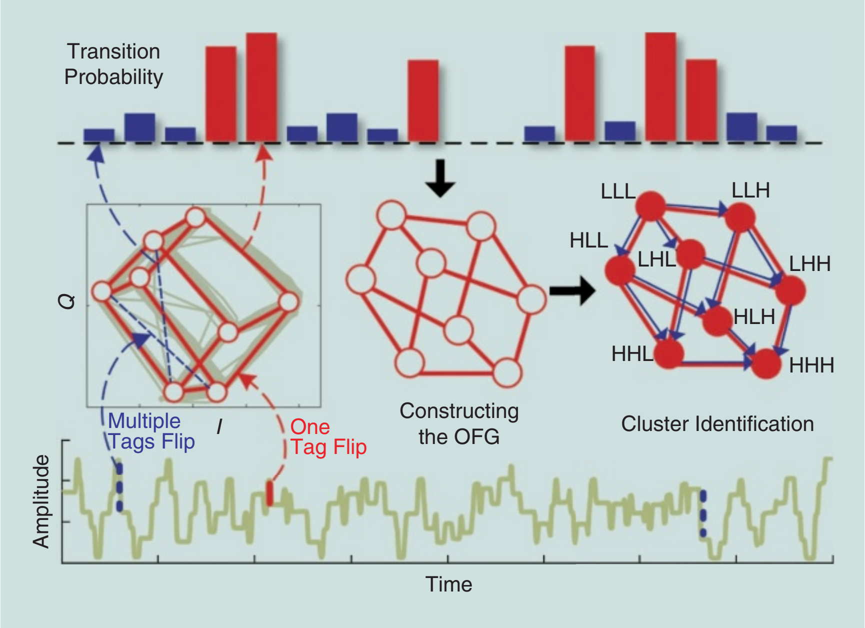


FIGURE 6. The workflow of FlipTracer. FlipTracer first constructs OFG in the IQ domain by selecting edges with large transition probabilities, then assigns each cluster to a parallel bits representation (three tags in this figure), and finally decodes the parallel bits with the OFG.

Harnessing channel state information for CTC

Compared to Wi-Fi, ZigBee has an orders-of-magnitude smaller maximum transmission power, and a much thinner channel bandwidth. These asymmetries of different communication standards make direct transmissions from ZigBee to Wi-Fi challenging. Our recent work ZigFi leverages channel state information (CSI), an indicator of Wi-Fi channel quality, to enable Wi-Fi to hear low-power ZigBee transmissions [23]. Figure 7 shows, when ZigBee transmissions interfere with Wi-Fi preambles, the changes of CSI amplitude offer a promising encoding space. In ZigFi, a Wi-Fi device decodes bytes by detecting the appearance and the absence of ZigBee signals at specific channels. By dedicatedly training time-series data classifiers, ZigFi can achieve a throughput of 215.9 bits/s, which is 18 times faster than the state of the art.

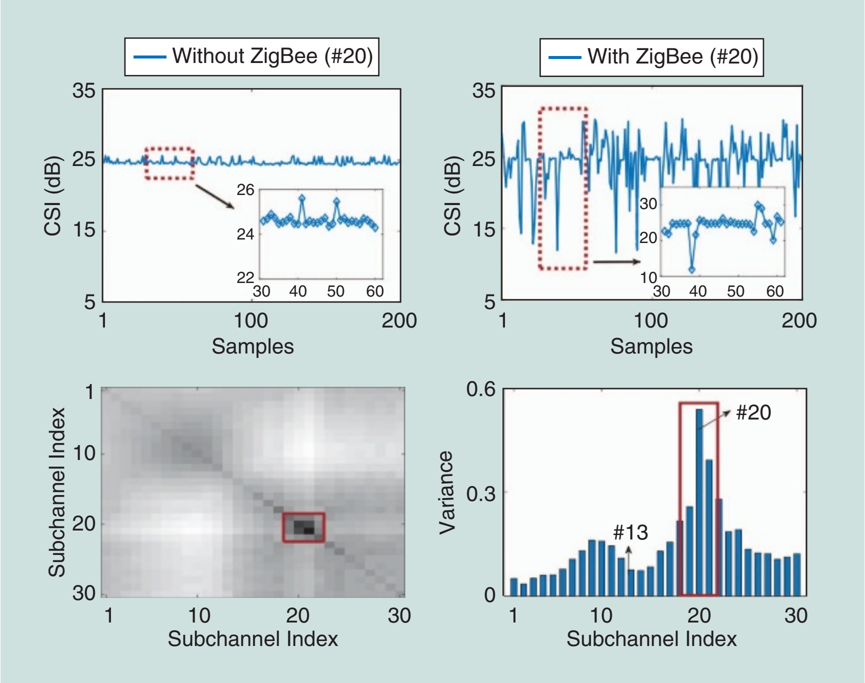


FIGURE 7. ZigFi: CTC from ZigBee to Wi-Fi. ZigFi discloses that ZigBee signals can interfere with Wi-Fi preambles and change the CSI pattern of specific subchannels, e.g., channel 20. Selected subchannels are used for CTC encoding.

Ongoing works

As mentioned previously, deep learning can provide effectiveness while edge computing can offer efficiency. A universal edge-computing architecture for real-time large-scale video analytics is desperately needed in Pavatar. Moreover, data sampling in Pavatar faces a severe problem of the category imbalance, since the anomaly states of synchronous compensators are very scarce. Therefore, modern learning techniques such as online imbalanced and hard sample mining for multisource time-series data can be further tailored to this problem.

Summary and conclusions

In this article, we surveyed and discussed the challenges and recent works toward digital twin, from sensing, networking, to analytics layer. We also presented Pavatar, a real-world IoT system for UHVCSs. We introduced our experience with Pavatar, and discussed the research issues as well as the future directions of industrial IoT. Industrial IoT is of great significance to the innovation of traditional industry. It envisions that we could automatically monitors and comprehensively simulates the factory throughout the entire life cycle, from production and manufacturing, operation to maintenance, to liberate the workforce and provide credible decision supports for industrial operations.

Acknowledgments

This work was jointly supported by National Key R&D Program of China 2017YFB1003000, the research fund from the State Grid of China on "Automatic Surveillance and Remote Diagnosis of Synchronous Compensator," the National Nature Science Foundation (NSF) of China numbers 61772306 and 61672320, NSF of China Key Projects 61432015 and 61632008, and China Postdoctoral Science Foundation number 2016M601034.

Authors

Yuan He (heyuan@tsinghua.edu.cn) received his B.E. degree in computer science and technology from the University of Science and Technology of China, Hefei, in 2003, his M.E. degree in computer software and theory from the Institute of Software, Chinese Academy of Sciences, Beijing, in 2006, and his Ph.D. degree in computer science and engineering from the Hong Kong University of Science and Technology, China, in 2010. He is currently an associate professor with the School of Software and Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China. His research interests include the Internet of Things, wireless networks, and mobile and ubiquitous computing. He is a Senior Member of the IEEE and a member of the Association for Computing Machinery.

Junchen Guo (juncguo@gmail.com) received his B.E. degree in computer science and technology from the University of Science and Technology of China, Hefei, in 2016. He is currently pursuing his Ph.D. degree in software engineering with the School of Software, Tsinghua University, Beijing, China. His research interests include the Internet of Things and mobile and ubiquitous computing.

Xiaolong Zheng (xiaolong@greenorbs.com) received his B.E. degree in software engineering from the Dalian University of Technology, China, in 2011 and his Ph.D. degree in computer science and engineering from the Hong Kong University of Science and Technology, China, in 2015. He is currently a postdoctoral research associate with the School of Software, Tsinghua University, Beijing, China. His research interests include the Internet of Things, wireless networks, and ubiquitous computing.

References

[1] P. Middleton, T. Tsai, M. Yamaji, A. Gupta, and D. Ruebe. (2017). Forecast: Internet of things – Endpoints and associated services. [Online]. Available: https://www.gartner.com/doc/3840665/forecast-internet-things–endpoints.

[2] Huawei Technologies Co. Ltd. NB-IoT. (2018). [Online]. Available: http://developer.huawei.com/ict/en/site-iot/product/nb-iot

[3] P. C. Evans and M. Annunziata, "Industrial internet: Pushing the boundaries general electric reports," General Electric Reports, 2012, pp. 1–37.

[4] L. Yang, Y. Li, Q. Lin, X.-Y. Li, and Y. Liu, "Making sense of mechanical vibration period with sub-millisecond accuracy using backscatter signals," in *Proc. ACM Mobile Computing and Networking*, New York, Oct. 3–7, 2016, pp. 16–28.

[5] J. Wang, J. Xiong, X. Chen, and D. Fang, "TagScan: Simultaneous target imaging and material identification with commodity RFID devices," in *Proc. ACM Mobile Computing and Networking*, Snowbird, UT, Oct. 16–20, 2017, pp. 288–300.

[6] Y. Zheng, Y. He, M. Jin, X. heng, and Y. Liu, "RED: RFID-based eccentricity detection for high-speed rotating machinery," in *Proc. IEEE Int. Conf. Computer Communications*, Honolulu, HI, Apr. 15–19, 2018.

[7] C. Gao, Y. Li, and X. Zhang, "LiveTag: Sensing human-object interaction through passive chipless WiFi tags," in *Proc. USENIX Networked Systems Design and Implementation*, Renton, WA, Apr. 9–11, 2018, pp. 533–546.

[8] C. Jiang, Y. He, X. Zheng, and Y. Liu, "Orientation-aware RFID tracking with centimeter-level accuracy," in *Proc. IEEE/ACM Information Processing in Sensor Networks*, Porto, Portugal, Apr. 11–13, 2018, pp. 290–301.

[9] F. Xiao, Z. Wang, N. Ye, R. Wang, and X.-Y. Li, "One more tag enables fine-grained RFID localization and tracking," *IEEE/ACM Trans. Netw.*, vol. 26, no. 1, pp. 161–174, 2018.

[10] Y. Ma, N. Selby, and F. Adib, "Minding the billions: Ultra-wideband localization for deployed RFID tags," in *Proc. ACM Mobile Computing and Networking*, Snowbird, UT, Oct. 16–20, 2017, pp. 248–260.

[11] G. Ananthanarayanan, P. Bahl, P. Bodík, K. Chintalapudi, M. Philipose, L. Ravindranath, and S. Sinha "Real-time video analytics: The killer app for edge computing," *IEEE Comput.*, vol. 50, no. 10, pp. 58–67, 2017.

[12] C. Long, Y. Cao, T. Jiang, and Q. Zhang, "Edge computing framework for cooperative video processing in multimedia IoT systems," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1126–1139, 2017.

[13] L. Cheng, J. Wang, Z. Cao, and Y. Liu, "ViTrack: Efficient tracking on the edge for commodity video surveillance systems," in *Proc. IEEE Int. Conf. Computer Communications*, Honolulu, HI, Apr. 15–19, 2018, pp. 1–9.

[14] Y. Guo, A. Yao, and Y. Chen, "Dynamic network surgery for efficient DNNs," in *Proc. Neural Information Processing Systems*, Barcelona, Spain, Dec. 5–10, 2016, pp. 1379–1387.

[15] S. Yao, Y. Zhao, A. Zhang, L. Su, and T. Abdelzaher, "DeepIoT: Compressing deep neural network structures for sensing systems with a compressor-critic framework," in *Proc. ACM Sensor Systems*, Delft, The Netherlands, Nov. 5–8, 2017, pp. 43–56.

[16] N. D. Lane, S. Bhattacharya, P. Georgiev, C. Forlivesi, L. Jiao, L. Qendro, and F. Kawsar, "DeepX: A software accelerator for low-power deep learning inference on mobile devices," in *Proc. IEEE/ACM Information Processing in Sensor Networks*, Vienna, Austria, Apr. 11–14, 2016, pp. 1–12.

[17] A. Mathur, N. D. Lane, S. Bhattacharya, A. Boran, C. Forlivesi, and F. Kawsar, "DeepEye: Resource efficient local execution of multiple deep vision models using wearable commodity hardware," in *Proc. ACM Mobile Systems, Applications, and Services*, Niagara Falls, NY, June 19–23, 2017, pp. 68–81.

[18] M. Jin, Y. He, X. Meng, Y. Zheng, D. Fang, and X. Chen, "FlipTracer: Practical parallel decoding for backscatter communication," in *Proc. ACM Mobile Computing and Networking*, Snowbird, UT, Oct. 16–20, 2017, pp. 275–287.

[19] J. Guo, Y. He, and X. Zheng. "Pangu: Towards a software-defined architecture for multi-function wireless sensor networks," in *Proc. IEEE Int. Conf. Parallel and Distributed Systems*, Shenzhen, China, Dec. 15–17, 2017, pp. 730–737.

[20] S. M. Kim and T. He, "FreeBee: Cross-technology communication via free side-channel," in *Proc. ACM Mobile Computing and Networking*, Paris, France, Sept. 7–11, 2015, pp. 317–330.

[21] X. Guo, X .Zheng, and Y. He. "WiZig: Cross-technology energy communication over a noisy channel," in *Proc. IEEE Int. Conf. Computer Communications*, Paris, France, May 1–4, 2017, pp. 1–9.

[22] Z. Li and T. He, "WEBee: Physical-layer cross-technology communication via emulation," in *Proc. ACM Mobile Computing and Networking*, Snowbird, UT, Oct. 16–20, 2017, pp. 2–14.

[23] X. Guo, Y. He, X. Zheng, L. Yu, and O. Gnawali, "ZigFi: Harnessing channel state information for cross-technology communication," in *Proc. IEEE Int. Conf. Computer Communications*, Honolulu, HI, Apr. 15–19, 2018, pp. 1–9.

[24] Z. Yin, Z. Li, S. M. Kim, and T. He, "Explicit channel coordination via cross-technology communication," in *Proceedings of ACM Mobile Systems, Applications, and Services*, Munich, Germany, June 10–15, 2018.

[25] Z. Yu, C. Jiang, Y. He, X. Zheng, and X. Guo, "Crocs: Cross-technology clock synchronization for Wifi and Zigbee," in *Proc. Embedded Wireless Systems and Networks*, Madrid, Spain, Feb. 14–16, 2018.

[26] X. Zheng, Y. He, and X. Guo, "StripComm: Interference-resilient cross-technology communication in coexisting environments," in *Proc. IEEE Int. Conf. Computer Communications*, Honolulu, HI, Apr. 15–19, 2018, pp. 1–9.

[27] A. Zhang, S. Song, J. Wang, and P. S. Yu, "Time series data cleaning: From anomaly detection to anomaly repairing," in *Proc. Very Large Data Bases Endowment*, Munich, Germany, Aug. 28–Sept. 1, 2017, pp. 1046–1057.

[28] D. Li, Y. Zhou, G. Hu, and C. J. Spanos, "Optimal sensor configuration and feature selection for AHU fault detection and diagnosis," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1369–1380, 2017.

[29] Y. Zhou, R. Arghandeh, and C. J. Spanos, "Partial knowledge data-driven event detection for power distribution networks," *IEEE Trans. Smart Grid*, 2017. doi: 10.1109/TSG.2017.2681962.

[30] Tsinghua University (2018). Pavatar project. [Online]. Available: http://tns.thss.tsinghua.edu.cn/sun/pavatar.html

Eva Arias-de-Reyna, Pau Closas,
Davide Dardari, and Petar M. Djurić

Crowd-Based Learning of Spatial Fields for the Internet of Things

From harvesting of data to inference

INTERNET OF THINGS—ISTOCKPHOTO.COM/IAREMENKO
CIRCUITS—IMAGE LICENSED BY INGRAM PUBLISHING

The knowledge of spatial distributions of physical quantities, such as radio-frequency (RF) interference, pollution, geomagnetic field magnitude, temperature, humidity, audio, and light intensity, will foster the development of new context-aware applications. For example, knowing the distribution of RF interference might significantly improve cognitive radio systems [1], [2]. Similarly, knowing the spatial variations of the geomagnetic field could support autonomous navigation of robots (including drones) in factories and/or hazardous scenarios [3]. Other examples are related to the estimation of temperature gradients, detection of sources of RF signals, or percentages of certain chemical components. As a result, people could get personalized health-related information based on their exposure to sources of risks (e.g., chemical or pollution). We refer to these spatial distributions of physical quantities as *spatial fields*. All of the aforementioned examples have in common that learning the spatial fields requires a large number of sensors (agents) surveying the area [4], [5].

A common way to sense environmental variables is the deployment of dedicated wireless sensor networks (WSNs), which continues to stimulate fertile research activities in the scientific community. Typical WSN applications are oriented to sense specific physical quantities (e.g., temperature) in well-defined areas [6], [7]. Unfortunately WSNs are generally characterized by significant constraints in terms of deployment cost, energy limitation, and the need for maintenance. These constraints prevent them from becoming scalable and therefore from being the ultimate solution for automated and distributed sensing of the physical world.

The expected pervasive diffusion of Internet of Things (IoT) devices (fixed and mobile) opens up a unique opportunity for a wide and massive sensing and mapping (i.e., georeferencing of physical quantities). In fact, the IoT constitutes a paradigm where a multitude of heterogeneous devices is able to sense the environment, process data, and actuate, thus creating the necessary infrastructure for cyberphysical systems. This infrastructure galvanizes technologies such as smart grids, smart homes, smart cities, and intelligent transportation [8], [9].

Digital Object Identifier 10.1109/MSP.2018.2840156
Date of publication: 28 August 2018

From the extensive variety of applications of the IoT, we are interested in those that will benefit from having a spatial coverage of a wide area due to a large number of agents navigating through it. For instance, this can be the case when devices are carried by people or autonomous agents (e.g., vehicles, robots, or drones) moving in outdoor- and indoor-populated environments like malls, stadiums, or crowded buildings. One can even imagine cities at large, if one considers much larger settings in size. Thanks to the widespread diffusion of IoT devices with heterogeneous sensors, the estimation of spatial physical fields is creating a new trend for next-generation sensor networks, referred to as *mobile crowdsensing networks* [10]–[13]. This is basically a zero-effort approach to automatically collect and process data. Recently, as an example, this concept has been proposed for zero-effort automatic mapping of environmental features using sensors already embedded in smartphones, such as magnetometers and Wi-Fi [14]–[17]. In such settings, the contribution of the agent to the sensing process is as simple as carrying the personal device in a pocket while the individual is moving around. Individuals are not even requested to be participatory, as the sensing process could run in the background during the normal operation of the device. In other words, agents are aware of the background sensing process, but they are not participatory in the sense that they are not requested to follow particular paths to make the learning process more effective. Thus, the sensing process is not an exclusive task, and it arises from the dynamic reality of humans or autonomous agents. The sensing process is a result of piggybacking on the capabilities of today's and future wireless personal devices. Including data generated by these devices will dramatically increase the amount of data for sensing and mapping purposes, with obvious benefits in terms of the resulting accuracy.

In this context, the IoT is the technological enabler for crowdsensing and learning of spatial fields. Interestingly, IoT devices are, in general, able to communicate among themselves, either directly or through a fusion node that can potentially be in the cloud. Thanks to communication capabilities, empirical data gathered by mobile agents (the crowd) can be collected and processed by learning algorithms located in the cloud. These algorithms exploit the correspondence between the position and the value of the physical quantity measured in that position to estimate the spatial field. As a consequence, positioning and spatial field estimation are intimately intertwined, as will be illustrated in the "Sensing and Positioning" section.

Crowdsourcing-based learning methods rely on the experience gained by previous agents. In principle, it is possible that, with crowd-based learning, one can perform optimal information fusion [10]. On the other hand, moving from the well-controlled conditions of WSN scenarios, where nodes are deployed in ad hoc known locations, to crowdsensing settings, where agents move around in an uncontrolled manner, entails a number of issues that need to be addressed. The methods rely on sharing through cloud mechanisms [18], but they can be of practical relevance in IoT applications only if their computational and memory requirements do not grow with the amount of collected data. Therefore, novel methodologies for multisensor data fusion and information processing are needed. They should guarantee efficient statistical representation of spatial fields and a computational complexity that does not depend on the number of measurements. Further, the algorithms need to be robust against irregular positioning and measurement errors.

Introduction

This article addresses the challenges and solutions of learning spatial fields for the IoT whose multitude of connected devices sense the fields. In many real-world scenarios, one may have measurements acquired by thousands of people or autonomous mobile agents interacting with each other and with things. The underlying idea is that each agent takes advantage of the measurements acquired by previous agents and, in turn, contributes to further improvement of the field estimates, which amounts to an indirect cooperative approach. More specifically, this article analyzes the main issues, techniques, and architectures for efficient crowd-based learning of spatial fields in the IoT. The nature of the problem suggests searching for solutions within the Bayesian framework, and this is what we have adopted. It is clear, however, that one may apply other methods including various types of data-driven or other non-Bayesian methods.

> At the beginning of each realization, the crowd-based learning method started without any knowledge about the true bias fields of the different anchors.

As previously pointed out, an even more challenging application of this concept is the joint positioning and spatial field learning in indoor environments, where agents aim at self-localization and, at the same time, learn the position-dependent parameters of the underlying observation models (represented as spatial fields). We put particular emphasis on this topic to show the great potential of crowd-based learning approaches.

In particular, in the section "Inference Methods for Learning Spatial Fields," we discuss the case of absence of specific and accurate models for the fields. Under this assumption, one approach to learning the spatial fields is to consider that the field is a sample from a Gaussian process (GP). In this article, we are interested in GPs and, particularly, their representations through linear combinations of orthogonal basis functions. This approach enables the development of inference algorithms whose complexity does not grow with the number of observations, which is necessary in the context of crowd-based learning.

In the "Sensing and Positioning" section, we discuss the issue of assuming perfect knowledge of the position where the data are sensed. This is almost never the case since the devices are typically positioned through the Global Navigation Satellite System (GNSS) or some other technology [19] and they provide position estimates with errors. It is therefore crucial to account for them in the process of learning spatial fields.

Otherwise, the obtained results will be unreliable and/or inaccurate. Two examples of joint tracking and crowd-based learning methods are discussed.

The "Use Case" section illustrates a use case related to the problem of indoor localization and tracking in the presence of biased ranging measurements caused by non-line-of-sight (NLOS) channel conditions, typical in time-based positioning systems, such as those based on the ultrawide-band (UWB) technology [20]. NLOS conditions might affect the position-dependent parameters of the observation model of the tracking algorithm and, thus, are treated as spatial fields to be estimated jointly with the agent position. Agents entering the area of interest take advantage of the knowledge inferred from data acquired by previous agents so that the expected tracking performance improves as the number of agents participating in the crowdsourcing grows, as shown in our experiments. An outlook of future directions of research is provided in the "Conclusions" section.

Inference methods for learning spatial fields

Suppose we want to estimate a static spatial field, which we denote with $f(\mathbf{x})$, where $\mathbf{x}$ contains location coordinates (the method can be modified to allow for the estimation of time-varying spatial fields). We make the initial estimate from T noisy observations $y_t = y(\mathbf{x}_t)$, acquired at known locations $\mathbf{x}_t, t = 1,2,\ldots,T$. The set of these observations and locations, $\mathcal{D} = \{(y_t,\mathbf{x}_t) | t = 1,2,\ldots,T\}$, represents a training set from which initial information about the spatial field $f(\mathbf{x})$ can be extracted. In particular, the objective is to make inference about the spatial field at locations without observations, i.e., not included in $\mathcal{D}$, and the corresponding uncertainty of the estimates. In the absence of specific and accurate models of the field, one approach is to assume that the field is a sample from a GP. We note that GPs represent a powerful and widely adopted machine-learning methodology [21].

The resulting nonparametric regression problem has a well-known solution, which unfortunately suffers from a computational complexity of the order of $O(T^3)$ [21]. It goes without saying that this solution is unaffordable in crowdsensing scenarios where T could grow to huge values. Several methods have been proposed to overcome this issue. For regular grids, fast solutions can be obtained through fast Fourier transform-based approaches [22] or by approximatively describing the GP through state-space models, making the complexity independent of T under certain conditions [23]. Other methods are referenced in [24].

Agents entering the area of interest take advantage of the knowledge inferred from data acquired by previous agents.

Most of the proposed solutions, however, are not computationally and memory efficient when applied to crowd-based learning, where all observations are not available simultaneously and might grow fast. Moreover, observations are usually obtained at random locations because the agents are generally not participatory, which entails that grid-based approaches cannot be applied in this context. On the other hand, we know that low-complexity incremental methods that update the field estimate once a new observation is acquired are preferable.

To tackle these issues, the authors in [25] and [26] propose a combined GP-state space method, whose complexity and memory requirements do not depend on the number of observations, in static and dynamic scenarios, respectively. This allows an efficient statistical characterization of the spatial field, which can be easily updated once new data become available, thus making it well suited for crowd-based learning applications. The main idea is represented in Figure 1 and summarized in the following.

Given an appropriate set of two-dimensional (2-D) orthogonal basis functions (e.g., the 2-D Fourier transform) in the area of interest, $\psi(\mathbf{x}) = [\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \ldots, \psi_J(\mathbf{x})]^\top$, the spatial field $f(\mathbf{x})$ is modeled by

$$f(\mathbf{x}) = \psi^\top(\mathbf{x})\mathbf{c}, \tag{1}$$

where $\mathbf{c}$ is a $J \times 1$ Gaussian vector of coefficients with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Thus, the GP is described by

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), \kappa(\mathbf{x},\mathbf{x}')), \tag{2}$$

where $m(\mathbf{x}) = \psi^\top(\mathbf{x})\boldsymbol{\mu}$ and $\kappa(\mathbf{x},\mathbf{x}') = \psi^\top(\mathbf{x})\boldsymbol{\Sigma}\psi(\mathbf{x}')$ are its mean and covariance, respectively.

The set of random coefficients $\mathbf{c}$ can be thought of as the state of a state-space model described by a hidden Markov process. The size of $\mathbf{c}$, J (dimension of the state-space) depends on the spatial variability of the physical field (spatial bandwidth). Thanks to (1), the problem of representing and estimating the spatial field $f(\mathbf{x})$ translates to characterizing the vector of coefficients $\mathbf{c}$, which does not depend on $\mathbf{x}$ and does not increase in size with the number of observations. Due to the Gaussian hypothesis, $\mathbf{c}$ is completely statistically characterized by the mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$.

The key to the crowd-based learning idea is to use all of the observed data collected until time t by all past agents for updating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. We denote the conditional vector of coefficients and its mean and covariance at time t by $\mathbf{c}_t$, $\boldsymbol{\mu}_t$, and $\boldsymbol{\Sigma}_t$, respectively. Note that, at time $t = 0$, $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ represent the a priori statistical knowledge about the GP. Each time a new noisy observation y_t is acquired, e.g., at position $\mathbf{x}_t$, it is used to update the characterization of the GP, $f(\mathbf{x})$, to $\hat{f}_t(\mathbf{x})$ given the observations $y_{1:t} = \{y_1, y_2, \ldots, y_t\}$ by using the model in (1) and (2). This can be accomplished by properly updating $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ of $\mathbf{c}_t$ from the previous conditional mean and covariance $\boldsymbol{\mu}_{t-1}$ and $\boldsymbol{\Sigma}_{t-1}$. Such an iterative learning process can be expressed in general as

$$(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) = \mathcal{L}[(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1}), y_t, \mathbf{x}_t], \tag{3}$$

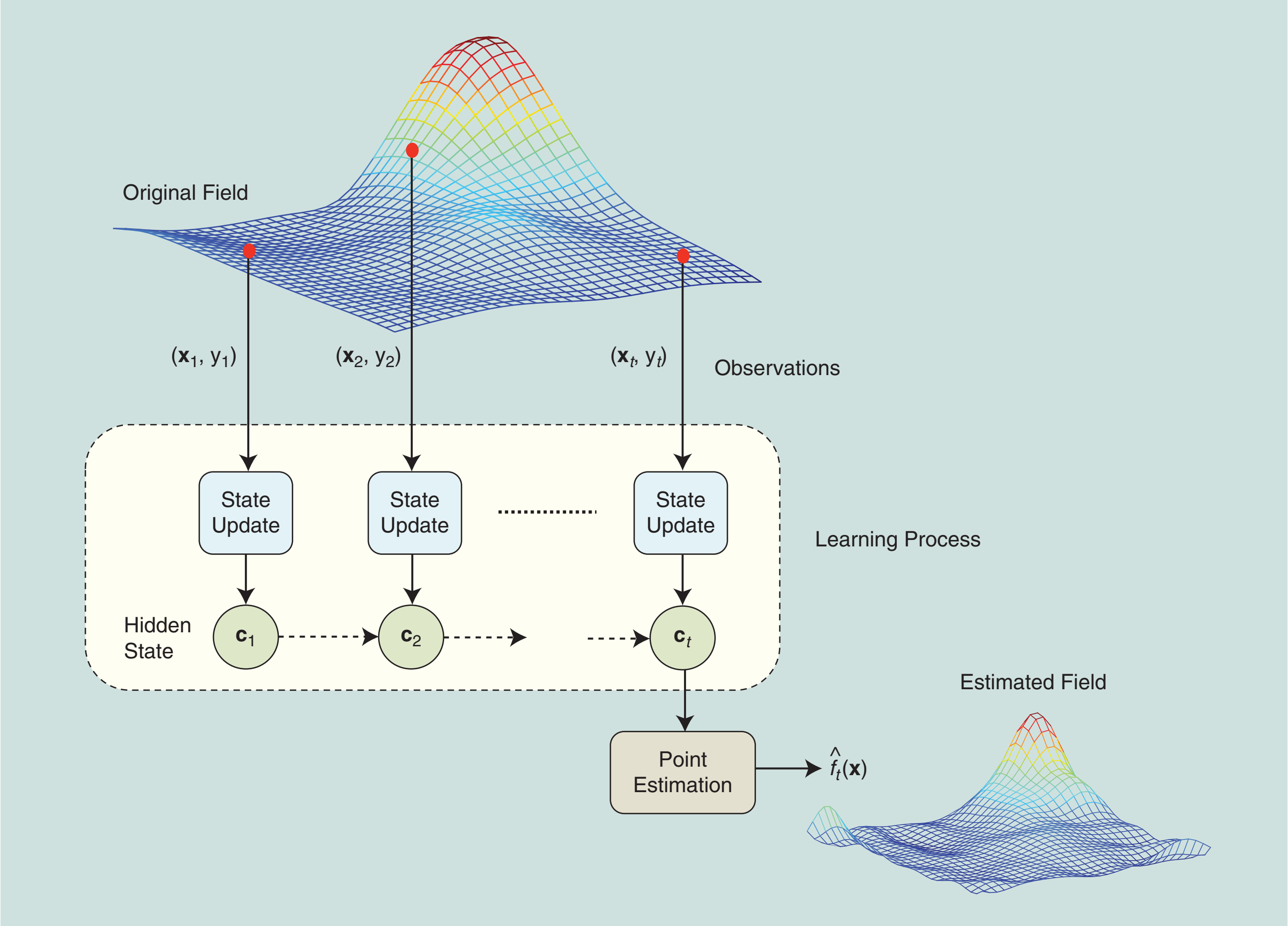


FIGURE 1. The hidden Markov process applied to crowd-based learning. The parameters $\mathbf{c}_t$ capture information about the spatial field $f(\mathbf{x})$ from data available up to time t, and $\hat{f}_t(\mathbf{x})$ is the estimate of the field at $\mathbf{x}$ after t measurements.

where $\mathcal{L}[\cdot,\cdot,\cdot]$ represents the learning algorithm that updates $(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$ to $(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ by considering the latest observation y_t and the position $\mathbf{x}_t$ at which it was collected. For instance, under the hypothesis of Gaussian observation noise and known $\mathbf{x}_t$, all the involved random variables are Gaussian, and the transformation (1) from the state vector to the spatial field is linear. Then the evolution of $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ in (3) can be computed efficiently through a Kalman filter algorithm with a complexity independent of the number of observations.

At any time t, one can compute a point estimate $\hat{f}_t(\mathbf{x})$ and/or the confidence interval of the GP at a certain position $\mathbf{x}$ conditioned on the history of observations by evaluating the maximum a posteriori (MAP) or minimum mean-square error (MMSE) estimate $\hat{\mathbf{c}}_t = \boldsymbol{\mu}_t$ of the coefficient vector $\mathbf{c}_t$ at time t and applying the transformation (1) (we denote the estimate of the vector $\mathbf{c}_t$ with $\hat{\mathbf{c}}_t$ to convey that the estimate is made at time t using $y_{1:t}$). Therefore, the updated mean and covariance $(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ represent the knowledge (sufficient statistics) about the GP acquired up to the current observation instant t. In this way, it is not necessary to keep in the memory all the past observations, whose number could grow to huge values. Instead, it suffices to store only $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$, whose sizes depend on J only.

Sensing and positioning

A common assumption in crowdsensing is that the agents sense the environment at perfectly known locations [5]. However, in the absence of this knowledge, their positions are typically estimated—through GNSS or other technologies [27]—and, thus, are somewhat uncertain. Moreover, other sources of error might be present such as the relative location between the agent's centroid and the mobile terminal. It is, therefore, crucial to account for this uncertainty in the process of learning the spatial fields; otherwise, the obtained results will be unreliable and/or inaccurate. Localization techniques, such as multilateration, fingerprinting, sources of opportunity, etc., rely on the availability of position-dependent physical measurements [e.g., time-of-arrival (TOA), received signal strength indicator] from which observations are derived. Regardless of the adopted localization technology, the main performance limitation is often imposed by model mismatches, i.e., the discrepancy between the reality and the applied models for characterizing the observations used in the localization process. Even if a model is accurate, the model parameters might depend on the agents' positions; hence, they can be treated as spatial fields to be estimated. In turn, estimates of the locations depend on the model parameters, i.e., on the (unknown) spatial fields. Such a “chicken and egg” problem is challenging and can be tackled through joint crowd-based learning and localization methods.

In what follows, we illustrate two possible approaches to estimate jointly the position and the field(s) characterizing the observation model parameter(s). For both approaches we suppose that a central unit or, more generally, the cloud keeps receiving the observations y_t (e.g., distance estimates) from all the agents in the space of interest. Based on these observations, this central unit has to estimate the position $\mathbf{x}_t$ of each agent and update the estimate of the coefficients $\mathbf{c}_t$ for the spatial field of interest.

Loose coupling approach

The first approach, which we refer to as *loose coupling*, was originally proposed in [28] and extended in [29]. According to the approach, each of the unknowns $\mathbf{x}_t$ and $\mathbf{c}_t$ is estimated by its own method. The two methods communicate with each other by exchanging sequentially their respective estimates (see Figure 2).

Specifically, at time t, the estimate of an agent’s position $\mathbf{x}_t$ is derived starting from the incoming position-dependent observation of that agent, y_t, as well as all the previous measurements through a tracking algorithm (the symbol y_t represents a scalar, but in general it can be a vector that contains multiple measurements):

$$p(\mathbf{x}_t | y_{1:t}) = \mathcal{P}[p(\mathbf{x}_{t-1} | y_{1:t-1}), y_t, \hat{f}_{t-1}(\cdot)], \qquad (4)$$

where $p(\mathbf{x}_t | y_{1:t})$ is the a posteriori probability density function (pdf) of $\mathbf{x}_t$, $\mathcal{P}[\cdot,\cdot,\cdot]$ denotes any iterative positioning/tracking algorithm able to provide the a posteriori pdf at time t as a function of the a posteriori pdf at time $t-1$, the new observations y_t and the estimate of the spatial field(s) $\hat{f}_{t-1}(\cdot)$, where the spatial field is a part of the observation model. Actually, there may be more than one spatial field of interest, as will be illustrated in the section “Use Case.” Such positioning/tracking algorithms can be derived using well-known Bayesian filtering tools that also allow for mobility models. From $p(\mathbf{x}_t | y_{1:t})$, an estimate $\hat{\mathbf{x}}_t$ of the position can easily be obtained, e.g., MAP, MMSE, or any other point estimate. More details can be found in [27].

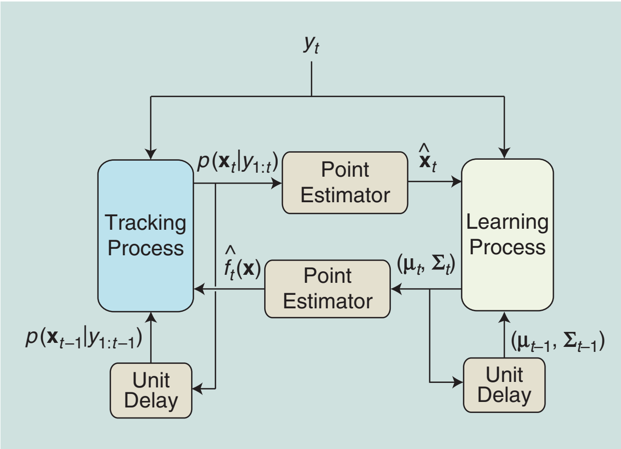


FIGURE 2. A schematic illustration of the loose coupling method.

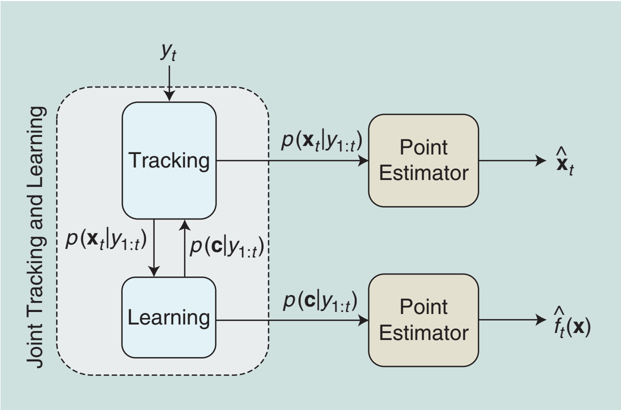


FIGURE 3. A schematic illustration of the tight coupling method.

The updated position estimate is then used as the input of the learning algorithm:

$$(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) = \mathcal{L}[(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1}), y_t, \hat{\mathbf{x}}_t], \qquad (5)$$

where $\mathcal{L}[\cdot,\cdot,\cdot]$, as in (3), updates $(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$ of $\mathbf{c}_t$ to $(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ by considering the latest observation y_t. Note that here, differently from (3), the estimated position $\hat{\mathbf{x}}_t$ is used instead of the true one, $\mathbf{x}_t$, which is not available. In addition, as detailed in the section “Use Case,” the observation used for training in the learning process might not be directly y_t, but a function of it and the estimated position $\hat{\mathbf{x}}_t$.

In the following step, the estimated field $\hat{f}_t(\cdot)$, obtained from $(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, is sent to the tracking algorithm, which obtains the estimate $\hat{\mathbf{x}}_{t+1}$. The algorithms continue to update and exchange estimates as new measurements keep coming.

The main advantage of this method is its relatively easy implementation because the tracking and learning tasks are separated. The learning algorithm amounts to a standard recursive least-squares-type method. For tracking, one can use any particular algorithm [particle filtering (PF), extended Kalman filtering, etc.]. The choice of filter would depend on various factors including the adopted mobility and observations models (typically nonlinear) and the allowed computational complexity of the tracking process.

Tight coupling approach

In the second approach, which we refer to as *tight coupling*, $\mathbf{x}_t$ and $\mathbf{c}$ are estimated by way of integrating out $\mathbf{c}$ while estimating $\mathbf{x}_t$, with $\mathbf{c}$ still being estimated (see Figure 3). At each recursion, the joint posterior pdf of the unknowns $\mathbf{x}_t$, $p(\mathbf{x}_t | y_{1:t})$, also known as *filtering pdf*, is calculated from the previous joint posterior and the new measurement y_t, or

$$p(\mathbf{x}_t | y_{1:t}) = \mathcal{T}[p(\mathbf{x}_{t-1} | y_{1:t-1}), y_t], \qquad (6)$$

where $\mathcal{T}[\cdot,\cdot]$ is an algorithm that performs the computation of the a posteriori pdf of $\mathbf{x}_t$. Unlike in (4), it appears that there is no use of $\mathbf{c}$ in the algorithm. However, this is not the case. More specifically, if we denote the posterior of $\mathbf{c}$ by $p(\mathbf{c} \mid y_{1:t})$ and according to the usual Markovian assumption for the underlying state-space model [27], the equations that describe how $p(\mathbf{x}_t \mid y_{1:t})$ and $p(\mathbf{c} \mid y_{1:t})$ are updated are expressed as follows:

$$p(\mathbf{x}_t \mid y_{1:t}) \propto \int p(y_t \mid \mathbf{x}_t, y_{1:t-1}) p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\mathbf{x}_{t-1} \mid y_{1:t-1}) \, \mathrm{d}\mathbf{x}_{t-1}, \tag{7}$$

$$p(\mathbf{c} \mid y_{1:t}) \propto \int p(y_t \mid \mathbf{x}_t, \mathbf{c}) p(\mathbf{x}_t \mid y_{1:t-1}, \mathbf{c}) p(\mathbf{c} \mid y_{1:t-1}) \, \mathrm{d}\mathbf{x}_t, \tag{8}$$

where $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ is the mobility model. These equations use two pdfs, $p(y_t \mid \mathbf{x}_t, y_{1:t-1})$ and $p(\mathbf{x}_t \mid y_{1:t-1}, \mathbf{c})$, defined by

$$p(y_t \mid \mathbf{x}_t, y_{1:t-1}) = \int p(y_t \mid \mathbf{x}_t, \mathbf{c}) p(\mathbf{c} \mid y_{1:t-1}) \, \mathrm{d}\mathbf{c}, \tag{9}$$

$$p(\mathbf{x}_t \mid y_{1:t-1}, \mathbf{c}) = \int p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\mathbf{x}_{t-1} \mid y_{1:t-1}, \mathbf{c}) \, \mathrm{d}\mathbf{x}_{t-1}. \tag{10}$$

Finally, the pdf $p(\mathbf{x}_{t-1} \mid y_{1:t-1}, \mathbf{c})$ that appears in (10) is obtained from

$$\begin{aligned} p(\mathbf{x}_{t-1} \mid y_{1:t-1}, \mathbf{c}) &\propto p(y_{t-1} \mid \mathbf{x}_{t-1}, \mathbf{c}) \\ &\times p(\mathbf{c} \mid y_{1:t-2}) \int p(\mathbf{x}_{t-1} \mid \mathbf{x}_{t-2}) p(\mathbf{x}_{t-2} \mid y_{1:t-2}) \, \mathrm{d}\mathbf{x}_{t-2}, \end{aligned} \tag{11}$$

where all the necessary pdfs in (11) are known from previous recursions.

Point estimates $\hat{\mathbf{x}}_t$ of the position or estimates of the predicted spatial field $\hat{f}_t(\cdot)$ can easily be obtained from $p(\mathbf{x}_t \mid y_{1:t})$ and $p(\mathbf{c} \mid y_{1:t-1})$ [as well as (1) and (2)]. As is evident from the previous equations, the design of the algorithm $\mathcal{T}[\cdot,\cdot]$ is more complex than that of the loose coupling approach.

Use case

We discuss a relevant use case in the context of joint crowd-based learning and localization. We consider agents navigating in a certain area whose behavior is modeled as a random walk, which is quite common in the literature in the absence of any other information on user (agent) behavior. Each user is recording TOA physical measurements with respect to fixed reference nodes. We call these nodes *anchor nodes* because we know their locations and we compute distance estimates with respect to them using TOA measurements. Such estimates could be subjected to unknown bias due to NLOS conditions that might characterize the channel between the agent and the ith anchor node according to the following observation model [20], [27]:

$$y^{[i]} = \left\| \mathbf{x} - \mathbf{x}_A^{[i]} \right\| + f^{[i]}(\mathbf{x}) + \nu^{[i]}, \tag{12}$$

where $\mathbf{x}_{\mathrm{A}}^{[i]}$ denotes the (known) position of the ith anchor, $\left\| \mathbf{x} - \mathbf{x}_{\mathrm{A}}^{[i]} \right\|$ is the true distance between anchor i and the agent, $\nu^{[i]}$ is a zero mean Gaussian perturbation, and $f^{[i]}(\mathbf{x})$ represents the spatial field characterizing the spatial behavior of the bias induced by the NLOS condition. Obviously, we have one spatial field for each anchor. It is worth highlighting that, in loose coupling, the distance observation $y^{[i]}$ cannot be used directly by the learning process because the latter needs bias (i.e., spatial field) observations that are not available. A possible solution is to derive a virtual bias observation $\tilde{y}^{[i]}$ from (12) using the estimated position $\hat{\mathbf{x}}$ instead of the true one, i.e., $\tilde{y}^{[i]} = y^{[i]} - \left\| \hat{\mathbf{x}} - \mathbf{x}_{\mathrm{A}}^{[i]} \right\|$.

While an agent crosses the area of interest, it takes advantage of the available estimated field obtained from measurements acquired by previous agents. In turn, the estimate of this field is updated by the measurements of this agent. Thereby, subsequent agents can also benefit by using the field for their own localization. In the "Introduction" section, we referred to this sharing of information among the users as *indirect cooperation*.

Now we present some numerical results for the NLOS use case validating the crowd-based approach discussed in this article. This set of representative experiments highlights the benefits of leveraging the crowd to learn the NLOS-induced bias field, ultimately improving the knowledge of the field at a reduced cost and without calibration requirements. We point out that, in the case of cooperative calibration with some users navigating along prescribed trajectories, performance could be improved.

We based the simulation of the observation values on real measurements taken in a typical office indoor environment with walls made of concrete; a floor plan is shown in Figure 4. We had four anchors, denoted in the figure as tx_i, $i = \{1,2,3,4\}$ (note that the figure also shows a fifth anchor, but its measurements were not used). Figure 4 also shows a set of 20 test locations where 1,500 range measurements were taken for each anchor using a commercial UWB radio operating in the 3.2–7.4-GHz band. The complete description of the measurement campaign can be found in [20].

For each anchor, the following procedure was carried out. For each of the 20 test locations we computed the mean and the standard deviation σ of the range measurements. Then we evaluated the bias as the difference between the mean range observation and the true distance. The value used in the simulations as true bias for any other location was obtained by interpolation from the set of bias values. Most test locations were in NLOS with respect to the anchors, thus making the localization and fields estimation processes quite challenging.

In the simulations, the users were entering in succession and moved randomly within a square area of side $L = 9$ m in Figure 4. Each user had a total of 50 measurements during the sojourn in the area, taken at intervals of 1 s. The simulated trajectories followed a random-walk model, with a noise covariance matrix equal to $\rho^2 \mathbf{I}$ with $\rho = 0.1$ and where $\mathbf{I}$ is the identity matrix, and the ranging measurements had errors with a standard deviation of $\sigma = 0.1$ m, in addition to the bias when in NLOS. Notice that, if another technology different from UWB

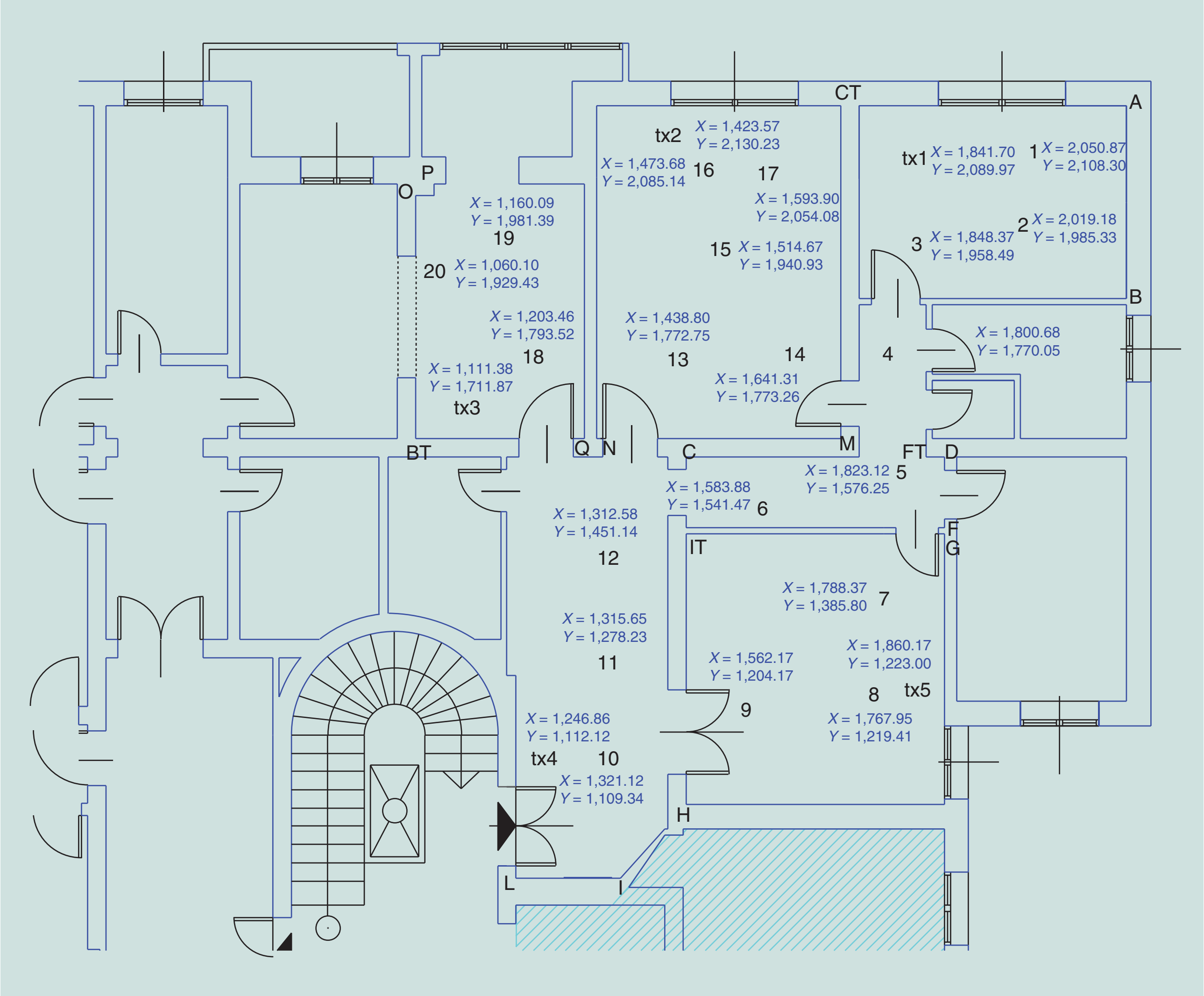


FIGURE 4. The floor plan of the office environment considered in the case study, based on the measurement campaign in [20]. There are four anchors of fixed known location and 20 test locations.

is used to collect $y^{[i]}$, the standard deviation of the measurements should be changed accordingly.

Figure 5(a) shows the corresponding spatial field $f^{[2]}(\mathbf{x})$ for anchor number two. The rest of the illustrations in Figure 5 correspond to results obtained by the loose coupling approach described in the previous section. As 2-D orthogonal basis functions $\psi(\mathbf{x})$ in the learning process, we used the exponentials of the 2-D Fourier series expansion of the periodical repetition of $f^{[i]}(\mathbf{x})$ with a period L in each dimension. They have been mapped into the corresponding real and imaginary components and truncated to $J = 578$ terms. Here, the 2-D Fourier exponentials have been taken as an example. In general, the choice of the basis functions is strictly dependent on the application and deserves further research. The tracking process was carried out by means of PF with 500 particles.

Figure 5(b)–(d) shows the estimated spatial field of anchor two, after 20, 50, and 200 users, respectively. More specifically, (b)–(d) shows the mean $m(\mathbf{x})$ of the GP representing the spatial field $f^{[2]}(\mathbf{x})$ at these instants. A progressive improvement of the field knowledge can be observed, which is visually apparent from the closer resemblance of the estimated fields to the true field in Figure 5(a) as the number of users grows.

Figure 6 shows the localization error (the distance between the true and estimated locations) as a function of the time step, for the 200th user of the same simulation of Figure 5. Besides the described technique with crowd-based learning, two more algorithms based on the same PF are included for comparison; one of them assumes an always-LOS condition while the other has perfect knowledge of the range bias value for each agent-anchor pair at each time instant. The latter is unrealistic and used as a benchmark, as it has the best performance, according to Figure 6. It can also be observed that the naive method assuming always-LOS has the worst performance. The crowd-based learning method achieves an improvement with respect to the method assuming LOS.

We conducted a Monte Carlo simulation by varying the set of user trajectories. Specifically, the simulation consisted

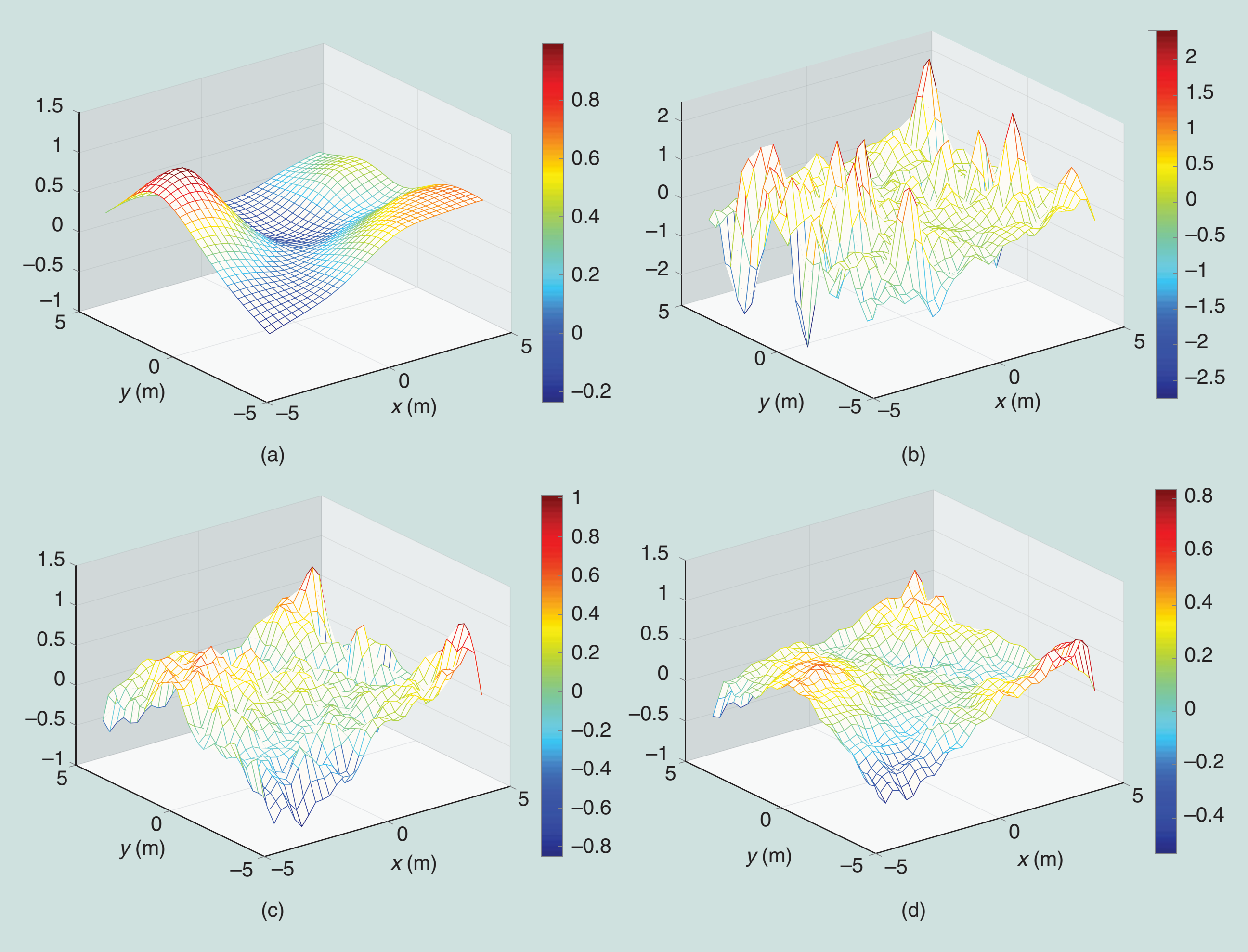


FIGURE 5. RF surveying can be achieved by crowd-based approaches, where knowledge is gained with agents navigating an area. For the described scenario and for anchor number two: (a) the true bias field and its estimation after (b) 20 users, (c) 50 users, and (d) 200 users.

of 500 Monte Carlo realizations where, at each realization, K users entered the area in sequence and followed a random path. The users were tracked and the field estimates of the different anchors were updated. At the beginning of each realization, the crowd-based learning method started without any knowledge about the true bias fields of the different anchors, i.e., the learning process was reset. The results in terms of (empirical) cumulative distribution function (CDF) of the localization error are shown in Figure 7. The curves encompass all the values of localization error for each of the 50 measurements per user and each of the 500 realizations. For the method with crowd-based learning, several curves are displayed, each of them encompassing the results for a specific set of users. The result improves as the number of users grows, first very quickly and later slowing down. Even in this challenging situation with typical NLOS range bias values prevailing and, in the absence of any calibration, the crowd-enhanced method achieves a notable improvement with respect to the simple method assuming always-LOS.

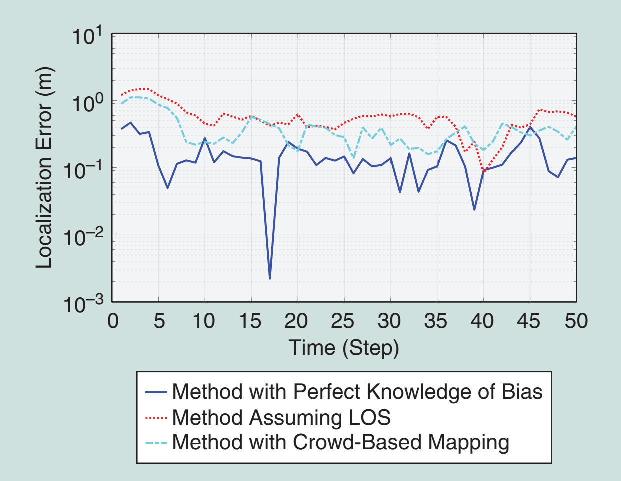


FIGURE 6. Positioning error as a function of time for the 200th user.

As an example, if one sets a target performance at 0.5 m with the described loose approach after 20 users have navigated in the area, 40% of the locations meet the target performance (covered) with respect to 30% assuming always-LOS. After 200 users, more than 60% of the area has the same target performance. The root-mean-square error for the crowd-based method, computed from the error values corresponding to the curve for users 191 to 200 is 0.72 m, an intermediate value between those of the method assuming LOS (0.88 m) and the benchmark with perfect knowledge (0.26 m).

Conclusions

We addressed the problem of estimating the spatial distribution of physical quantities (spatial fields) by taking an advantage of the pervasive diffusion of IoT mobile devices equipped with sensors collecting measurements related to the spatial field at different locations. This crowd-based learning scenario, where the knowledge about the field is refined as new agents enter the area, poses several challenges mainly caused by the random and uncertain position of the agents and the unbounded growth of the amount of collected data. We showed how these issues can be tackled within the framework of signal processing by illustrating a couple of methods for efficient joint learning and positioning (in terms of memory and computational burden). To demonstrate the potential of these methods, one of them was applied to the problem of indoor positioning in the presence of NLOS using real measurements in which the range bias value was modeled as a spatial field to be jointly estimated with the location of the agent.

> **Information fusion by multiple agents accessing the cloud in an asynchronous manner and distributed learning by the agents are potentially fertile directions for future research investigations.**

Several other topics, such as energy consumption minimization, privacy-preserving schemes, and the incentivization of agents to participate in the crowdsensing process [30], deserve further investigation. The choice of basis functions in representing the spatial fields and the dimension of the spanned space is also important.

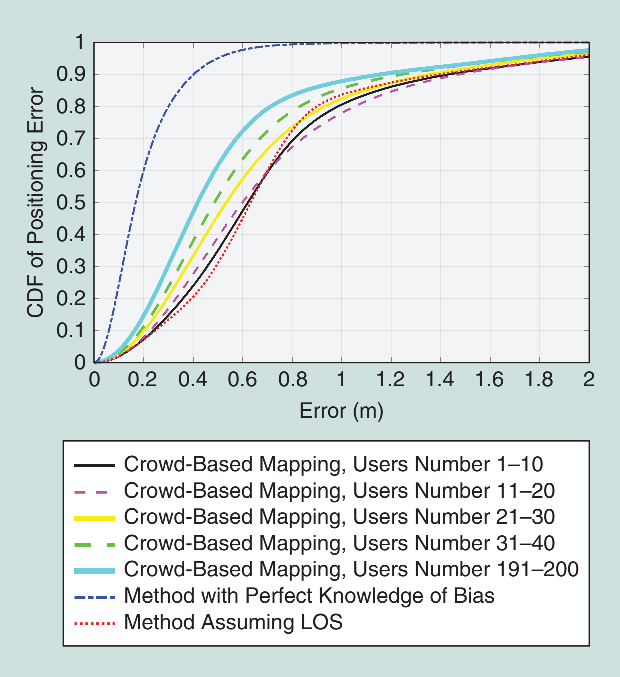


FIGURE 7. The CDF of localization error for the three methods considered.

Information fusion by multiple agents accessing the cloud in an asynchronous manner and distributed learning by the agents are potentially fertile directions for future research investigations. For example, we will commonly have two or more agents that will be tracked simultaneously, and there is more than one way to fuse the information extracted from the measurements about the spatial field. In the case of loose coupling, the fusion is less challenging because the system needs to keep track basically of the mean and covariance of **c** for each spatial field. The update of these statistics can take place after a new measurement is received (from any agent), and the tracking algorithm will always be fed with the latest statistics. Further, the updates can be asynchronous. By contrast, the tight coupling approach provides interesting challenges because implementations of fusion after every received measurement are not easy. The alternative is to fuse the information after the agent leaves the area. Yet another set of questions about fusion arises when the agents do not transmit their measurements to the central unit and instead, given the estimates of the spatial fields from the central unit, they track themselves and at the same time continue to improve the estimates of the spatial fields. At some point in time, before leaving the area, they report their estimates of all the spatial fields to the central unit, which now has to fuse them with the existing information.

Last but not least, the number of computations needed to implement the tracking of the agents and the update of the spatial fields can also be an issue. When most of the computations take place away from the agents, this is not so critical. However, when the agents employ apps for self-tracking as suggested previously, then it is essential that the required need for computing power is minimized. So it is expected that there will be research in methods that minimize computational costs while maintaining guaranteed accuracy of self-localization.

Acknowledgments

Eva Arias-de-Reyna was supported by the Spanish Ministry of Economy and Competitiveness and the Spanish National Research Agency under grant TEC2016-78434-C3-2-R (AEI/FEDER, EU), by the European Union (FEDER) and by Junta de Andalucía (TIC-155). Davide Dardari was supported by the European H2020 project XCycle (grant 635975). Petar M. Djurić was supported by the National Science Foundation under Awards CCF-1618999 and CNS-1642965.

Authors

Eva Arias-de-Reyna (earias@us.es) received her M.S. and Ph.D. degrees in telecommunication engineering from the Universidad de Sevilla, Spain, in 2001 and 2007, respectively, where she is currently an associate professor with the Department of Signal Theory and Communications, which she joined in 2002 after working in industry for a year. Her current research interests include machine learning, localization techniques, ultrawide-band technology, and communication receiver design. She is a Senior Member of the IEEE.

Pau Closas (closas@northeastern.edu) received his M.S. and Ph.D. degrees in electrical engineering from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 2003 and 2009, respectively. He also received an M.S. degree in advanced mathematics from UPC in 2014. He is an assistant professor in the Electrical and Computer Engineering Department at Northeastern University, Boston, Massachusetts. His primary areas of interest include statistical and robust signal processing, Bayesian inference, and stochastic filtering. He is the recipient of the 2014 EURASIP Best Ph.D. Thesis Award, the 2014 Duran Farell Award for Technology Research, and the 2016 ION Early Achievements Award.

Davide Dardari (davide.dardari@unibo.it) received his laurea degree (summa cum laude) in electronic engineering and his Ph.D. degree in electronic engineering and computer science from the University of Bologna, Italy, in 1993 and 1998, respectively. He is an associate professor at the University of Bologna, Italy. Since 2005, he has been a research affiliate at the Massachusetts Institute of Technology, Cambridge. His research interests are in wireless communications, localization techniques, and distributed signal processing. He received the IEEE Aerospace and Electronic Systems Society's M. Barry Carlton Award (2011) and the IEEE Communications Society Fred W. Ellersick Prize (2012). He was the chair for the Radio Communications Committee of the IEEE Communication Society and a Distinguished Lecturer (2018–2019). He served as an editor of *IEEE Transactions on Wireless Communications* from 2006 to 2012. He is a Senior Member of the IEEE.

Petar M. Djuric′ (petar.djuric@stonybrook.edu) received his B.S. and M.S. degrees in electrical engineering from the University of Belgrade, and his Ph.D. degree in electrical engineering from the University of Rhode Island. He is a SUNY Distinguished Professor in the Department of Electrical and Computer Engineering at Stony Brook University, New York. His research has been in various areas of signal and information processing with an emphasis on Bayesian methods. He received the IEEE Signal Processing Magazine Best Paper Award in 2007 and the EURASIP Technical Achievement Award in 2012. He is a Fellow of the IEEE and EURASIP.

References

[1] R. Di Taranto, S. Muppirisetty, R. Raulefs, D. Slock, T. Svensson, and H. Wymeersch, "Location-aware communications for 5G networks: How location information can improve scalability, latency, and robustness of 5G," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 102–112, Nov. 2014.

[2] J. Lunden, V. Koivunen, and H. V. Poor, "Spectrum exploration and exploitation for cognitive radio: Recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 123–140, May 2015.

[3] W. Storms, J. Shockley, and J. Raquet, "Magnetic field navigation in an indoor environment," in *Proc. IEEE Conf. Ubiquitous Positioning Indoor Navigation and Location Based Service*, 2010, pp. 1–10.

[4] I. Nevat, G. W. Peters, and I. B. Collings, "Random field reconstruction with quantization in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 6020–6033, Dec. 2013.

[5] J. Unnikrishnan and M. Vetterli, "Sampling and reconstruction of spatial fields using mobile sensors," *IEEE Trans. Signal Process.*, vol. 61, no. 9, pp. 2328–2340, May 2013.

[6] R. Verdone, D. Dardari, G. Mazzini, and A. Conti, *Wireless Sensor and Actuator Networks: Technologies, Analysis and Design*. London: Elsevier, 2008.

[7] D. Dardari, A. Conti, C. Buratti, and R. Verdone, "Mathematical evaluation of environmental monitoring estimation error through energy-efficient wireless sensor networks," *IEEE Trans. Mobile Comput.*, vol. 6, no. 7, pp. 790–802, July 2007.

[8] J. Shi, J. Wan, H. Yan, and H. Suo, "A survey of cyber-physical systems," in *Proc. Int. Conf. Wireless Communications and Signal Processing*, Nov. 2011, pp. 1–6.

[9] C. X. Mavromoustakis, G. Mastorakis, and J. M. Batalla, *Internet of Things (IoT) in 5G Mobile Technologies*. Cham, Switzerland: Springer Int. Publishing, 2016.

[10] H. Ma, D. Zhao, and P. Yuan, "Opportunities in mobile crowd sensing," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 29–35, 2014.

[11] B. Guo, Z. Wang, Z. Yu, Y. Wang, N. Y. Yen, R. Huang, and X. Zhou, "Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm," *ACM Comput. Surv.*, vol. 48, no. 1, pp. 7:1–7:31, Aug. 2015.

[12] M. Zappatore, A. Longo, and M. A. Bochicchio, "Crowd-sensing our smart cities: a platform for noise monitoring and acoustic urban planning," *J. Commun. Softw. Syst.*, vol. 13, no. 2, pp. 53–67, June 2017.

[13] J. Liu, H. Shen, and X. Zhang, "A survey of mobile crowdsensing techniques: A critical component for the Internet of Things," in *Proc. 25th Int. Conf. Computer Communication and Networks*, Aug. 2016, pp. 1–6.

[14] A. Rai, K. K. Chintalapudi, V. N. Padmanabhan, and R. Sen, "Zee: Zero-effort crowdsourcing for indoor localization," in *Proc. ACM 18th Annu. Int. Conf. Mobile Computing and Networking*, New York, 2012, pp. 293–304.

[15] M. Alzantot and M. Youssef, "Crowdinside: Automatic construction of indoor floorplans," in *Proc. ACM 20th Int. Conf. Advances Geographic Information Systems*, New York, NY, 2012, pp. 99–108.

[16] T. Higuchi, H. Yamaguchi, and T. Higashino, "Context-supported local crowd mapping via collaborative sensing with mobile phones," *Pervasive Mobile Comput.*, vol. 13, pp. 26–51,

[17] I. Koukoutsidis, "Estimating spatial averages of environmental parameters based on mobile crowdsensing," *ACM Trans. Sens. Netw.*, vol. 14, no. 1, pp. 2:1–2:26, Dec. 2017.

[18] E. S. Lohan, J. Torres-Sospedra, H. Leppäkoski, P. Richter, Z. Peng, and J. Huerta, "Wi-Fi crowdsourced fingerprinting dataset for indoor positioning," *Data*, vol. 2, no. 4, p. 32, 2017.

[19] P. Closas, M. Luise, J. Ávila-Rodriguez, C. Hegarty, and J. Lee, "Advances in signal processing for GNSSs [From the Guest Editors]," *IEEE Signal Process. Mag.*, vol. 34, no. 5, pp. 12–15, 2017.

[20] D. Dardari, A. Conti, J. Lien, and M. Z. Win, "The effect of cooperation on localization systems using UWB experimental data," *EURASIP J. Adv. Signal Process*, vol. 2008, p. 12, Feb. 2008.

[21] C. E. Rasmussen, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[22] J. Fritz, I. Neuweiler, and W. Nowak, "Application of FFT-based algorithms for large-scale universal Kriging problems," *Math. Geosci.*, vol. 41, no. 5, pp. 509–533, 2009.

[23] S. Särkkä, A. Solin, and J. Hartikainen, "Spatio-temporal learning via infinite-dimensional Bayesian filtering and smoothing," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 51–61, 2013.

[24] A. Solin and S. Särkkä, "Hilbert space methods for reduced-rank Gaussian process regression," arXiv Preprint, arXiv:1401.5508, 2014.

[25] D. Dardari, A. Arpino, F. Guidi, and R. Naldi, "A combined GP-state space method for efficient crowd mapping," in *Proc. IEEE Workshop Advances Network Localization and Navigation*, London, June 2015, pp. 761–765.

[26] D. Dardari, G. Pasolini, and F. Zabini, "An efficient method for physical fields mapping through crowdsensing," to be published.

[27] D. Dardari, P. Closas, and P. M. Djurić, "Indoor tracking: Theory, methods, and technologies," *IEEE Trans. Veh. Technol.*, vol. 64, no. 4, pp. 1263–1278, Apr. 2015.

[28] E. Arias-de-Reyna, D. Dardari, P. Closas, and P. M. Djurić, "Enhanced indoor localization through crowd sensing," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Mar. 2017, pp. 2487–2491.

[29] E. Arias-de-Reyna, D. Dardari, P. Closas, and P. M. Djurić, "Estimation of spatial fields of NLOS/LOS conditions for improved localization in indoor environments," in *Proc. IEEE Statistical Signal Processing Workshop*, June 2018, to be published.

[30] L. G. Jaimes, I. J. Vergara-Laurens, and A. Raij, "A survey of incentive techniques for mobile crowd sensing," *IEEE Internet Things J.*, vol. 2, no. 5, pp. 370–380, Oct. 2015.

Petros Spachos, Ioannis Papapanagiotou,
and Konstantinos N. Plataniotis

Microlocation for Smart Buildings in the Era of the Internet of Things

A survey of technologies, techniques, and approaches

INTERNET OF THINGS—ISTOCKPHOTO.COM/IAREMENKO
CIRCUITS—IMAGE LICENSED BY INGRAM PUBLISHING

Microlocation plays a key role in the transformation of traditional buildings into smart infrastructure. Microlocation is the process of locating any entity with a very high accuracy, possibly in centimeters. Such technologies require high detection accuracy, energy efficiency, wide reception range, low cost, and availability. In this article, we provide insights into various microlocation-enabling technologies, techniques, and services and discuss how they can accelerate the incorporation of the Internet of Things (IoT) in smart buildings. We cover the challenges and examine some signal processing filtering techniques such that microlocation-enabling technologies and services can be thoroughly integrated with an IoT-equipped smart building. An experiment with Bluetooth Low-Energy (BLE) beacons used for microlocation is also presented.

Overview of microlocation

The interconnectedness of all things is continuously expanding. The aim is to have every individual interconnected with his or her surroundings, whether it be at home, at work, or in public spaces. Some of these services might include but are not limited to indoor mapping and personalized environment changes, such as lighting and temperature settings, as well as directed advertisement. For these systems to perform, it is essential to have reliable hardware and accurate data. Outdoor localization technologies, such as the Global Positioning System (GPS), do not work indoors due to the physical barriers that block the signal and do not provide location data accurate enough for microlocation. Current solutions use received signal strength indication (RSSI) to determine position. A variety of solutions that use RSSI have been proposed to provide location services for indoor environments, though each solution presents its own drawbacks. Multiple technologies and techniques have been adapted to provide indoor location information, all of which attempt to overcome the noise and dynamics of a changing indoor environment.

A promising approach includes the effective use of the plethora of IoT devices that are available on the market. BLE

Digital Object Identifier 10.1109/MSP.2018.2846804
Date of publication: 28 August 2018

1053-5888/18©2018IEEE

beacons, usually referred to as *beacons*, are a promising candidate to improve indoor localization accuracy. They are small Bluetooth transmitters designed to attract attention to a specific location. As in many IoT-based networks, the performance of such networks relies on the network lifespan and accuracy. BLE beacons are a cheap, simple, and very scalable means of implementing indoor localization services. In recent years, BLE technology has grown in popularity, and much more research has been developed in using it for indoor localization [1]–[3]. The fundamental operation of these beacons for localization purposes is based on RSSI techniques, where the received RSSI value is translated into a distance by using a best curve fit signal propagation model. BLE beacon protocols, such as iBeacon [4] and Eddystone [5], provide the necessary information and configuration capabilities for microlocation. Along with the low power consumption of BLE, beacon devices are easily deployed and require low maintenance, hence their scalability for any complex indoor environment.

Intrinsic to any wireless technology, BLE beacons are highly susceptible to noise and interference. To overcome the effects of noise and dynamic changes in the physical environment, many methods devised around advanced positioning algorithms and filtering techniques have been adapted to beacon-based systems to improve the accuracy obtained in using RSSI localization techniques, as shown in Figure 1. Some of the most common filter implementations are Kalman filters, as detailed in [6]. Kalman filtering has also been examined in the context of indoor localization [7]. These filters provide a reasonably accurate state estimation and can be adjusted for changes is environmental/process noise. Other filters, such as particle filters (PFs), are also used. PFs are highly accurate but at the cost of greater computational complexity, hence the need for a client-server-based model, as outlined in [2] and [8]. Positioning algorithms can also have an effect on beacon accuracy. The work presented in [9] implements the K-nearest neighbor algorithm to calculate the position of the user. The experiments showed an average error of 1 m. Other algorithms, such as the pedestrian dead-reckoning approach, have been implemented with BLE beacons [10]. In these experiments, the integration of smartphone sensors for data regarding step detection, step direction, and walking length are combined with beacon calibration zones to provide a more accurate position. All techniques may provide different accuracy results and may behave differently depending on the environment, so it is important to note the characteristics of each tested environment when deciding on what technique to implement.

In this article, we survey available wireless technologies for microlocation systems in a smart building. Then we discuss signal processing techniques and characteristics that can be used to improve microlocation performance, along with filtering approaches. We focus on the use of BLE beacons, and, through an experiment, we discuss how they can enhance microlocation.

Smart buildings with IoT technologies

The IoT revolution has brought a swarm of continuously interconnected and sensor-packed devices opening a vast number of opportunities in equipping existing infrastructures. The IoT has enabled applications that transform facilities to intelligent spaces able to critically affect and improve the productivity and life quality of the occupants. Reducing energy costs and detecting and building knowledge based on human patterns as well as improving the human–building interaction are only some cases in point.

The Institute for Building Efficiency [11] defines smart buildings as buildings that can provide low-cost services, such as air conditioning, heating, ventilation, illumination, security, sanitation, and various other services, to tenants without adversely affecting the environment. This requires the collaboration of multiple sensors that form a building's IoT ecosystem. The basic motive behind the construction of smart buildings is to provide the highest level of comfort and efficiency. At the same time, the interconnection of the automation systems can assist with disaster management and provide emergency services. The collaboration of the fire system with the air conditioning system, e.g., can create an environment where a fire will not expand to the rest of the building.

To that end, indoor-focused location-based services (LBSs) are the fundamental components for providing a tenant-to-building interaction. LBSs provide the ability to efficiently track occupants in real time. They either attempt to estimate the user's two-dimensional (2-D) coordinates, which is referred to as *microlocation*, or they assign the user in the locality of certain points of interest, which is known as *proximity sensing.*

The integration of smart buildings with the IoT creates a number of challenges. A smart building with an IoT ecosystem requires three main components: the sensors, the integration, and the actuators. The sensors must be connected to a reliable, highly available network that optimally can self-diagnose and heal. Integration is probably the part where innovation is now taking place. It consists of some software that would receive the input from the sensors, process and

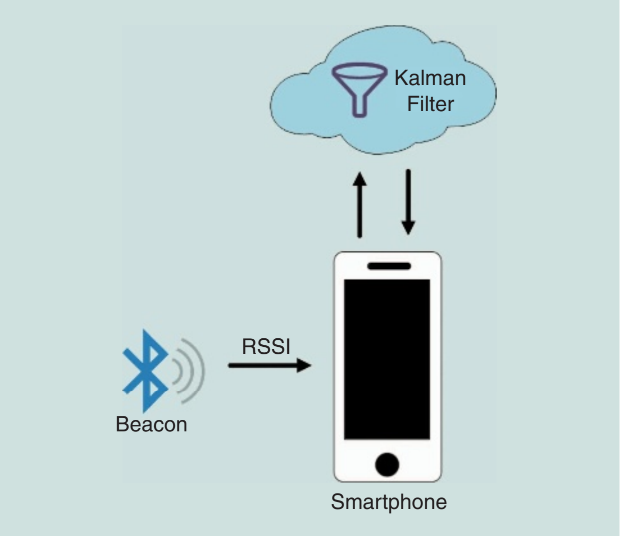


FIGURE 1. The microlocation system using the Kalman filter.

analyze, and provide some actuator as a service to the tenants, e.g., unlocking a door, switching on the TV, calling the elevator, or configuring the room temperature based on needs.

Overview of microlocation systems

Wireless technologies

Microlocation systems can leverage existing wireless infrastructure for microlocation to minimize the cost or may require a specific wireless deployment [1]. By *wireless technologies* we refer both to high-frequency technologies as well as low frequency. The most common high-frequency wireless technologies that have been used in a microlocation deployment are, e.g., Wi-Fi [12], Zigbee [13], radio-frequency identification (RFID) [14], and Bluetooth [15]. However, low-frequency technologies like the ones based on physical light have also seen some research and commercial use [16]. Light fidelity (Li-Fi), e.g., is one of the wireless technologies in the form of visible light communication (VLC) technology. These technologies have been used successfully in the past for indoor location and navigation, and their popularity among IoT devices makes them an ideal solution for microlocation as well. There are also technologies such as Wi-Fi HaLow [17], BLE version 5.0 [15], and LoRaWAN [18], which are specifically designed for IoT devices.

IEEE 802.11, Wi-Fi

The IEEE 802.11 standard [12], commonly known as *Wi-Fi*, is among the most popular technologies used for localization when GPS is inadequate. The great distribution of access points and signal availability at an indoor environment make it easy to collect the received signals from various access points and calculate the location of the receiver. The indoor transmission range can vary from 3.3 m with a bandwidth of 6.7 Gbit/s (IEEE 802.11ad), up to 70 m with a bandwidth of 600 Mbit/s (IEEE 802.11n), and it can operate in 2.4, 5, and 60 GHz.

Wi-Fi networks are deployed for communication; hence, data rate and connectivity are important, whereas localization is not their priority. Also, Wi-Fi networks are designed for a plethora of devices, from smartphones and laptops to phablets and smartwatches. This is a tradeoff for microlocation techniques. The availability of Wi-Fi signals and Wi-Fi-enabled devices is an advantage for microlocation as the number of portable devices and potential reference points for localization increases. Advanced signal processing techniques can be used to improve the quality of the Wi-Fi signals for localization. At the same time, there is no need for extra hardware deployment with Wi-Fi technology.

However, IoT devices have unique characteristics, such as size and limited energy resources, that are not taken into consideration for general Wi-Fi technology. As the number of these devices increases, the 2.4- and 5-GHz channels become overcrowded, whereas the interference increases with a drop in the network capacity. Unfortunately, traditional Wi-Fi was not originally designed to tackle these interference issues and the increasing capacity in dense environments. To fill this gap, the Wi-Fi Alliance announced the Wi-Fi HaLow (IEEE 802.11ah).

IEEE 802.11ah, Wi-Fi HaLow

Wi-Fi HaLow [17] was designed to enable connectivity to a variety of new power-efficient use cases in smart homes, smart cities, and connected vehicles and supporting the concept of the IoT in general. It extends Wi-Fi into the 900-MHz band to enable the low power connectivity that is necessary for IoT devices. The transmission range is twice the range of Wi-Fi, and it increases the signal robustness in challenging environments, such as complex indoor environments with lots of furniture and walls. It can operate in multiple transmission modes from low rates, starting from 150 and up to 347 kilobit/s.

The ability to operate in the low-power, high-transmission range and low propagation loss make Wi-Fi HaLow a good candidate for microlocation with IoT devices. However, it is relatively new in comparison with other technologies (published in 2017); hence, it is not widely available, and it will be a while before we see HaLow clients and infrastructure devices. This delays the experimentation that is necessary before deciding if it is suitable for microlocation.

Zigbee

Zigbee is a high-level communication protocol known for its simplicity, low power usage, and secure networking [13]. It is based on the IEEE 802.15.4 standard, which defines the operating point of wireless personal area networks (WPANs) with low-data-rate antennas. They are able to control the flow of information and prevent any loss of data by using carrier-sense multiple access with collision avoidance. Devices using Zigbee are designed with additional features, such as link quality and energy detection, that allow for measurements, such as the RSSI, to be easily determined. Zigbee is commonly used for localization in wireless sensor networks due to its low power requirements. Among IoT devices, though, it is not popular due to the extra hardware that is needed.

Bluetooth

Bluetooth is another wireless technology for exchanging data over short distances [15]. The IEEE standardized Bluetooth as IEEE 802.15.1 but no longer maintains the standard, which is managed by the Bluetooth Special Interest Group (SIG). According to the SIG, Bluetooth is all about proximity, not about exact location. Bluetooth was not intended to offer a pinned location like GPS. However, it is known as a *geofence* or *microfence solution*, which makes it an indoor proximity solution, not an indoor positioning solution.

Introduced by the Bluetooth SIG in 2010, BLE was designed for applications that do not require large amounts of data transfer, reducing the power consumption and cost of devices. Microlocation and indoor mapping have been linked to

Bluetooth and to the BLE-based iBeacon promoted by Apple [4]. Large-scale indoor positioning systems based on iBeacons have been implemented and applied in practice.

Similar to Zigbee, BLE is a technology used in WPANs. The low power consumption of BLE has led to a number of new devices in the IoT. BLE 4.0 can reach 25 Mbit/s at a distance of 60 m. Applications using BLE have greatly increased during the past couple of years. A number of new devices have been developed, in such fields as health care [19], sports, fitness, security, and home entertainment. One device that has been created is known as a *beacon*. Beacons are small, inexpensive devices that contain only a central processing unit, a radio, and batteries.

Bluetooth 5.0 [15] is the competitor of Wi-Fi HaLow in the IoT domain. It is claimed to have twice the speed of the previous version, four times longer transmission range, and exchange data eight times faster. The simplicity and popularity among IoT devices are advantages of Bluetooth for microlocation. The small size of beacons and their low cost with the energy efficiency of the BLE and the extended lifespan that it can provide can be used to enhance microlocation in a complex environment without interfering with other wireless infrastructures. For disadvantages, even though the security of BLE is good, it is even better on Wi-Fi.

RFID

RFID devices were primarily designed for data transfer and storage [14]. There is a need for an RFID reader that can communicate with RFID tags. There are two types of RFIDs. The active RFIDs operate in the ultrahigh frequency and microwave frequency ranges. They need to be connected to a local power source while they transmit their ID periodically up to 100 m. Passive RFIDs, however, operate without battery but within 1–2-m transmission range.

In the IoT era, RFID is not a promising solution for microlocation. Its accuracy is not high enough, and it is not available on many portable devices.

LoRaWAN

LoRaWAN is a long-range, low-power-consumption technology used in the development of personal wide area networks [18]. Originally developed by the LoRa Alliance, the LoRaWAN protocol transmits at a lower frequency of 915 MHz. The benefit of using a lower frequency is that the smaller wavelength allows for a greater distance that the signal can reach. Due to that, it can pass through walls and obstacles without issue. It is also no longer as easily susceptible to noise because it does not interfere with any devices transmitting on the 2.4-GHz band.

The disadvantage of using such a low frequency is a reduction in the data rate that can be sent between transmitting devices. For microlocation, this is not an issue, as the nodes are not transmitting large amounts of information. Due to the 915-MHz band being unlicensed, it is free for anyone to use for his or her personal networking needs.

For devices that are moving at high speed in a large area, LoRa might be a candidate for localization with the IoT. Unfortunately, in the short range, LoRa performance does not overcome the high cost and the extra equipment that are needed to set up a LoRa node.

Li-Fi

Li-Fi is a VLC technology [20]. VLC is a subset of optical wireless communication, which uses light-emitting diodes (LEDs) as a medium to enable high-speed communication. Data are transmitted by modulating the intensity of LED light at nanosecond intervals, too quick to be detected by the human eye.

Table 1 summarizes the specifications of each wireless technology along with the advantages and disadvantages of usage for microlocation.

Radio signal features for microlocation

As the wireless signal propagates from the sender to the receiver, there are signal characteristics that can be used for the

Table 1. The wireless technologies for microlocation.

Technology	Throughput	Transmission Range	Power Consumption	Advantages	Disadvantages
IEEE 802.11ac	3.5 Gbit/s	35 m	Moderate	Available in many environments	Prone to noise and interference
IEEE 802.11ad	6.7 Gbit/s	3.3 m			
IEEE 802.11ah	347 Mbit/s	1 km	Low	Wide reception range	Not widely available
Zigbee	250 kbit/s	75 m	Low	Easy to set up	Extra hardware
BLE v4.0	25 Mbit/s	60 m	Low	High throughput	Prone to interference
BLE v5.0	50 Mbit/s	240 m			
RFID active	1,067	100 m	Low	Low power	Low accuracy
RFID passive	1,067	2 m			
LoRaWAN	50 kbit/s	15 km	Extremely low	Wide range	Extra hardware
Li-Fi	1 Gbit/s	10 m	Low	Dense environments	Low range

localization of one of the communicating devices. There are four main signal features that can be used for localization.

RSSI

RSSI is one of the most commonly used characteristics for indoor localization [1]. It is based on measuring the power present in a received signal from a client device to an access point. As radio waves propagate according to the inverse-square law, the distance can be approximated based on the relationship between transmitted and received signal strength, as long as no other errors contribute to faulty results. The combination of this information with a propagation model can help to determine the distance between the client device and the access points. Lateration-based methods are commonly used along with RSSI to estimate the location of the client.

It can be assumed that the more access points, the more information can be collected, and hence the accuracy can be increased. This, however, works also as a tradeoff. An increase of the access points will also increase the interference between different signals. A key challenge in wireless localization systems is that the range measurements are often associated with errors. Although RSSI techniques are among the cheapest and easiest methods to implement, the disadvantage is that RSSI does not provide very good accuracy, with a median of 2–4 m. This is mainly because the RSSI measurements tend to fluctuate according to environmental changes or multipath fading, events that are common in indoor environments.

Angle of arrival

Angle of arrival (AoA) is another characteristic that can be used for localization. It tries to estimate the direction of the signal propagation, i.e., the angle from which the signal arrives at a receiver. AoA is typically achieved by using an array of antennas. The line connecting two reference points may be used as an internal reference. The spatial separation of antennas leads to differences in arrival times, amplitudes, and phases.

Time of arrival

In time of arrival (ToA) (also known as *time of flight*), the distance between the sender and receiver of a signal can be determined using the measured signal propagation time and the known signal velocity. ToA is the amount of time a signal takes to propagate from transmitter to receiver. The signal propagation rate is constant and known; hence, the travel time of a signal can be used to directly calculate distance. This is the technique used by GPS.

The accuracy of the ToA-based methods often suffers from massive multipath conditions in indoor localization, which is caused by the reflection and diffraction of the RF signal from objects (e.g., interior wall, doors, or furniture) in the environment. However, it is possible to reduce the effect of multipath by applying temporal or spatial sparsity-based techniques.

Time difference of arrival

The time difference of arrival (TDoA) is the ToA of a specific signal at physically separate receiving stations with precisely synchronized time references. TDoA measures the difference in ToA at two different receivers. Three or more TDoA measurements can be used to locate a device with hyperbolic lateration.

Although TDoA sounds similar to ToA, there is a difference. In ToA, the absolute time at a base station is used. In TDoA, the measured time difference between departing from one and arriving at the other station is used.

Indoor positioning techniques

Proximity detection

Proximity detection techniques, shown in Figure 2(a), are based on the proximity of the mobile device to previously known locations. These techniques determine the position of an object based on closeness to a reference in the physical space. When the mobile device receives the signal from a reference point, then the device should be within the coverage range of the reference point, i.e., in close proximity to the reference point. Proximity detection does not provide the location in the form of coordinates but rather in the form of sets of possible locations.

This method is also based on the premise that the reference point has a limited range. For simplicity, it is common to assume that the range of a wireless infrastructure would be well represented by a circle of given radius r. Then, the result of the proximity detection would be located inside this circle. For several circles, one can limit the possible location to the intersection of the different circles.

Lateration

Lateration is the process of estimating the location of a mobile device's given distance measurements to a set of points with a known location, shown in Figure 2(b). Lateration-based methods use the distance measurements from multiple reference points to compute the position of a receiver. Trilateration is a commonly used technique to calculate the estimated client device position relative to the known position of three access points. It uses the distance from the three reference points to estimate the location and track the position of the receiver when the receiver is moving within the three points. Given the distance to an anchor, it is known that the node must be along the circumference of a circle centered at the anchor and a radius equal to the node–anchor distance. In 2-D space, at least three noncollinear anchors are needed; in three-dimensional space, at least four noncoplanar anchors are needed.

Angulation

Angulation-based positioning techniques can be used to employ the AoA of a wireless signal and determine the position of a receiver, as shown in Figure 2(c). A commonly used approach is triangulation, where the location of a point is determined by forming triangles to it from known points. In

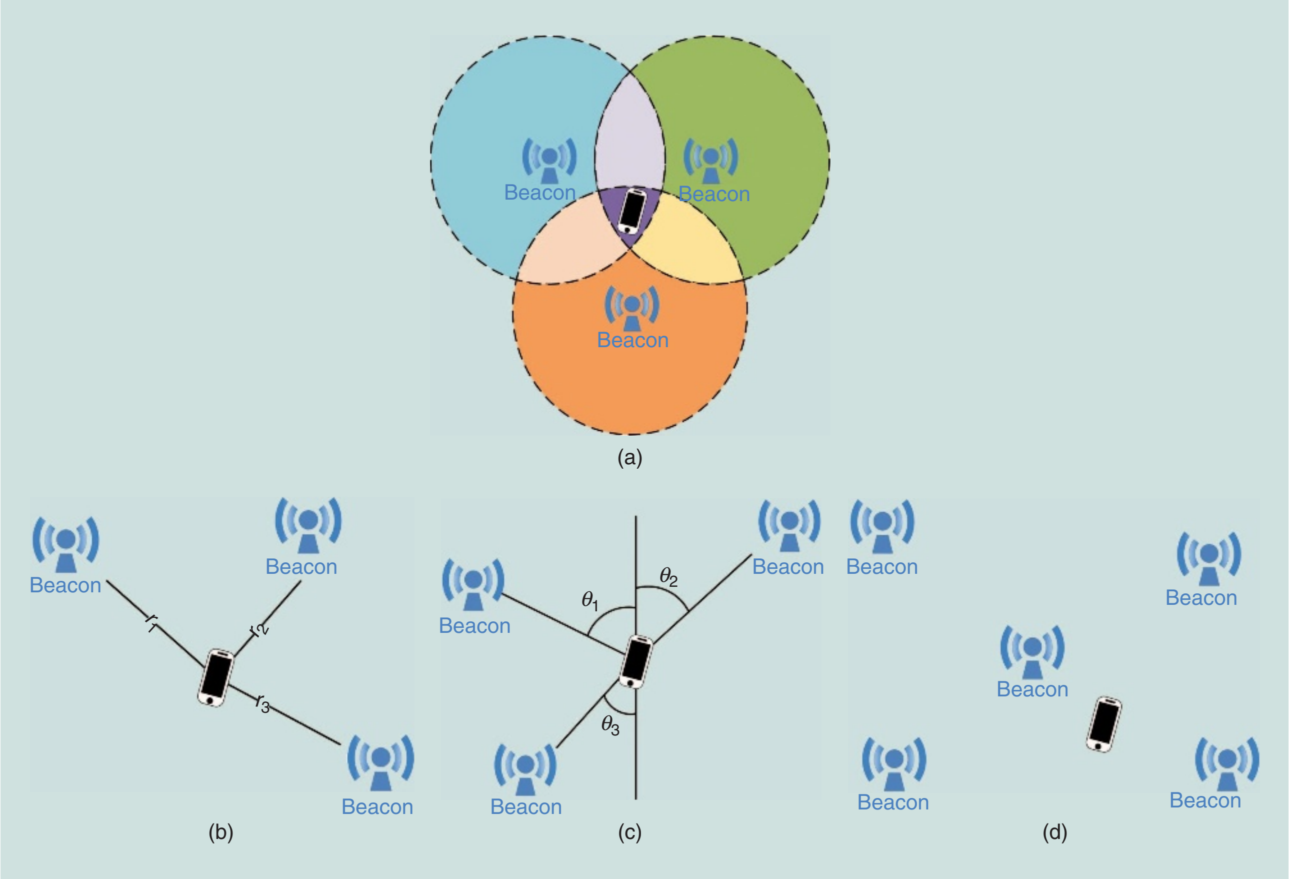


FIGURE 2. The localization techniques: (a) proximity, (b) lateration, (c) angulation, and (d) fingerprinting.

triangulation, a known baseline can be used to find the location relative to two anchor transmitters. It uses the geometric properties of triangles to estimate the location and relies on angle (bearing) measurements. It requires a minimum of two bearing lines and the locations of anchor nodes or the distance between them for 2-D space.

Fingerprinting

Fingerprinting techniques are based on the reproducibility of patterns of measurable variables, shown in Figure 2(d). Traditional fingerprinting records the signal strength from several access points and stores them in a database, along with the known coordinates of the client device in an offline phase. Then, during the localization phase, the current vectors at an unknown location are compared to those in the database, and the closest match is returned as the estimated user location.

Fingerprinting has the advantage that it does not require any assumption regarding the nature of the propagation environment. It just creates a model environment based on the training data. At the same time, this can be a disadvantage. Any change of the environment, such as adding or removing furniture or access points, requires an update to the model.

Localization metrics

To evaluate the performance of a localization system, accuracy and precision are used. Accuracy measures the deviation of the estimated location from the truth, whereas precision measures the deviation of location estimates from each other for the same location. A system with high accuracy can be used for an application that focuses on long-term localization determination, and the errors cancel out over time. A system with high precision can be used to find the proximity between devices, but it is hard to use for localization.

Improve accuracy through signal processing filtering techniques

There are a number of signal processing filtering techniques that are used for indoor localization. In the following, we summarize two: Kalman filtering and dynamic Kalman filtering.

Indoor localization model

We model the indoor localization problem as posed by Arulampalam et al. [21]. Extended versions as applied in BLE can also be found in [7]. Because we seek to estimate the user position/state under a set of measurements obtained in a typical noisy indoor environment, Bayesian filtering is an attractive approach for such problems. However, Bayesian filtering requires the following two models.

1) *System model*: A system model describes the variation of the state (user position in our case) with time. The system model relates the position vector y_i with the process noise m_i and previous state.

2) *Measurement model*: A measurement model relates the noisy measurements (RSSI for PF and the user position for extended Kalman filtering) with the state/position.

We construct the posterior probability density function (pdf) describing the state from all available information, including the measurements from the reference nodes (beacons in our case). The pdf is considered as the complete solution to the state estimation problem because it contains all of the required information. The problem involves recursively estimating the user state/position as we receive measurements from the beacon. Therefore, we require a recursive filter. Recursive filters consist of the prediction and update stage in which the state is predicted and then updated once the measurements are available. The presence of noise in indoor settings affects the position calculation, so the pdf is usually distorted. The obtained measurements in the update state are used to modify the prediction pdf using Bayes' theorem.

Mathematically, state y_i at time i is a function of the state at time step $(i–1)$ as well as the process noise m_{i-1} [22], as described in (1):

$$y_i = f_i(y_{i-1}, m_{i-1}). \tag{1}$$

The nonlinear function $f_i: \Re^{n_y} \times \Re^{n_m} \rightarrow \Re^{n_y}$ (as indoor localization is a nonlinear problem) relates the previous state y_{i-1} and process noise m_{i-1} with the current state *yi* as described by Arulampalam [21]. The sequence $\{m_i, i \in \aleph\}$ represents an independent and identically distributed (i.i.d.) process noise sequence. The integer n_y represents the state noise vector, and n_m represent the process noise vector. The set of natural numbers is represented by $\aleph$. The measurement model relates the obtained measurement x_i to the state y and measurement noise *n* at time *i* [22] as given in (2):

$$x_i = h_i(y_i, n_i). \tag{2}$$

The mapping function $h_i: \Re^{n_y} \times \Re^{n_n} \rightarrow \Re^{n_x}$ can be either linear or nonlinear. Functions f_i and h_i rely on the laws of motion/physics. The sequence $\{n_i, i \in \aleph\}$ is a measurement noise sequence that is i.i.d. The integers n_x and n_n represent the measurement and measurement noise vectors dimension, respectively.

Recursively calculating the pdf $p(y_i|x_{1:i})$ allows us to continuously calculate the belief in the state y_i at any particular time instance i in the presence of noisy measurements. The initial pdf $p(y_o|x_0)$ is assumed to be equivalent to the state vector's prior $p(y_0)$ [21]. We assume that the prior is available. The available information is enough to calculate the pdf $p(y_i|x_{1:i})$ recursively in the prediction and update stages. In the prediction stage, if the pdf $p(y_{i-1}|x_{1:i-1})$ is available, we can use the Chapman–Kolmogorov equation given in (3) to obtain the prior pdf of the state at any time instance *i*:

$$p(y_i|x_{1:i-1}) = \int p(y_i|y_{i-1})p(y_{k-1}|x_{1:i-1})dy_{i-1}. \tag{3}$$

At any time instance *i*, we collect the observations x_i from the sensors to update the prior using Bayes' rule given in (4) [21]. The denominator in (4) is explained in (5):

$$p(y_i|x_{1:i}) = \frac{p(x_i|y_i)p(y_i|x_{1:i-1})}{p(x_i|x_{i-1})}, \tag{4}$$

$$p(x_i|x_{i-1}) = \int p(x_i|y_i)p(y_i|x_{i-1})dy_i. \tag{5}$$

The collected measurements x_i in the update stage are then used to update the prior density, resulting in the required current state's posterior density. Recursively updating the system using (3) and (4) results in an optimal Bayesian solution. However, analytically, it is not possible to obtain the recursive propagation of posterior probability density as done in (3) and (4). Therefore, a number of different algorithms, including PF, Kalman filter, and extended Kalman filter, are used to obtain a solution.

Kalman filter

The Kalman-filter-based RSSI smoother is based on the work of Guvenc [23]. The state x_i, which in our case consists of RSSI and rate of change of RSSI, at time *i* is a function of the state at time *i*–1 and the process noise w_{i-1}, which is given mathematically by (6). The obtained RSSI measurements z_i at instant *i* from the iBeacons is a function of the state at *i*–1 and the measurement noise v_i as given by (7), as described in Arulampalam [21]:

$$x_i = f(x_{i-1}, w_{i-1}), \tag{6}$$

$$z_i = h(x_{i-1}, v_i). \tag{7}$$

The traditional Bayesian-based approach consists of the prediction and update stage, as described by Guvenc [23], and is given as follows:

1) prediction stage:

$$p(x_i|z_{1:i-1}) = \int p(x_i|x_{i-1})p(x_{i-1}|z_{1:i-1})dx_{i-1}. \tag{8}$$

2) update stage:

$$p(x_i|z_{1:i}) = \frac{p(z_i|x_i)p(x_i|z_{1:i-1})}{p(z_i|z_{1:i-1})}, \tag{9}$$

where

$$p(z_i|z_{1:i-1}) = \int p(z_i|x_i)p(x_i|z_{1:i-1})dx_i. \tag{10}$$

We assume that both the process noise and measurement noise are Gaussian and the functions *f* and *h* in (6) and (7) are linear. As a result of the linearity assumption, we can apply a Kalman filter because it is the optimal linear filter.

Due to the aforementioned assumptions, (6) and (7) can be rewritten as described by Guvenc [23]:

$$x_i = Fx_{i-1} + w_i, \tag{11}$$

$$z_i = Hx_i + v_i, \tag{12}$$

where $w_i \sim N(0, Q)$ and $v_i \sim N(0, R)$. Table 2 lists the parameters of a Kalman filter. The prediction and update stages for the Kalman filter as described by Guvenc [23] are

1) prediction stage:

$$\hat{x}_{\bar{i}} = F\hat{x}_i, \tag{13}$$

$$P_{\bar{i}} = FP_{i-1}F^T + Q. \tag{14}$$

2) update stage:

$$K_i = P_{\bar{i}}H^T(HP_{\bar{i}}H^T + R)^{-1}, \tag{15}$$

$$\hat{x}_i = \hat{x}_{\bar{i}} + K_i(z_i - H\hat{x}_{\bar{i}}), \tag{16}$$

$$P_i = (I - K_iH)P_{\bar{i}}. \tag{17}$$

The higher the Kalman gain, the higher will be the influence of the measurements on the state. The prediction and update steps are recursive in nature.

For the purpose of filtering the RSSI values, we use a state vector x_i that consists of the RSSI value y_i and the rate of change of RSSI Δy_{i-1} as follows: $x_i = \begin{bmatrix} y_i \\ \Delta y_i \end{bmatrix}$.

Depending on the environment, Δy_i signifies how drastically RSSI value fluctuates. The higher the noise in the environment, the higher will be the fluctuation. The current value of RSSI y_i is assumed to be the previous RSSI y_{i-1} plus the change Δy_i and process noise w_i^y. Hence (11) can be written as

$$\begin{bmatrix} y_i \\ \Delta y_i \end{bmatrix} = \begin{bmatrix} 1 & \delta t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{i-1} \\ \Delta y_{i-1} \end{bmatrix} + \begin{bmatrix} w_i^y \\ w_i^{\Delta y} \end{bmatrix}, \tag{18}$$

which means that the state transition matrix F is given by

$$F = \begin{bmatrix} 1 & \delta t \\ 0 & 1 \end{bmatrix}.$$

The parameter δt is to be adjusted as per the variation in RSSI, which depends on the environment. For our set of experiments, δt was taken as 0.2 (using trial and error). Similarly, (12) can be rewritten as

$$[z_i] = [1 \ \ 0] \begin{bmatrix} y_i \\ \Delta y_i \end{bmatrix} + [v_i^y]. \tag{19}$$

The observation matrix H is given by

$$H = [1 \ \ 0].$$

Parameters P, Q, and R used in the experiments were obtained using trial and error and are as follows:

$$P = 100\mathbf{I}_{22}, \ Q = 0.001\mathbf{I}_{22}, \ R = [0.10].$$

The Kalman filter, once calibrated, effectively smooths the RSSI values. The smoothed RSSI values were then input into the path-loss model to obtain distances between the iBeacons and the user, and the user's proximity to the beacon was classified in any of the aforementioned zones.

Table 2. The Kalman filter parameter notation.

Symbol	Meaning
x	State vector
z	Measurement/observation vector
F	State transition matrix
P	State vector estimate covariance or error covariance
Q	Process noise covariance
R	Measurement noise covariance
H	Observation matrix
K	Kalman gain
w	Process noise
v	Measurement noise

Dynamic Kalman

A dynamic variation of the Kalman filter computes Q as the variance of a set number of previously collected RSSI values to make up for real-world process noise changes. It is continuously recalculated at each iteration of reading in the next RSSI value.

Different set sizes of recorded RSSI values can be used to find the ideal number of values to use in this calculation. It can be inferred that as the array size increases, the accuracy increases as well, up to an array size of n. After a size of n, any increase of the size leads to a decrease of the accuracy. The optimal n can be found through experimentation, whereas the increase of the size can lead to waste of resources without any increase in the accuracy. At the same time, a decrease of the size below n does not give sufficient information to the system to increase its prediction accuracy.

The set of RSSI values is stored in an array list of data type double. Algorithm 1 illustrates the procedure. It adds entries to each index in increasing order starting from index 0. When an entry is deleted, all entries ahead get pushed down one index value. At the start of each iteration, the algorithm checks the size of the array; once it reaches the desired size n, it removes the oldest entry (index 0) and adds in the newest measurement.

When developing the dynamic noise component of the Kalman filter, it is essential to find the ideal number of previously obtained RSSI values to maintain, for calculation purposes. This is because the size of this set will have a direct impact on the performance of the filter.

Algorithm 1. Maintain RSSI set.

```
if $RSSIArray.size() == n$ then
    remove $RSSIArray[0]$.
$lastIndex \leftarrow RSSIArray.size()$
$RSSIArray[lastIndex] \leftarrow newRSSI$
```

BLE beacon technology

BLE beacons

BLE beacons are small wireless transmitters that broadcast their identifier to nearby electronic devices, such as smartphones, wearables, and other IoT devices. An analogy of the way beacons work is with the operation of a lighthouse. The lighthouse represents a known location that can be uniquely identified by its light. All of the ships that can see the light know about the existence of the lighthouse. However, the lighthouse neither communicates with the ships, nor does it know how many ships see its light or how many other lighthouses are in the area. Similarly, every beacon is sending out a radio signal to inform all of the radio-enabled devices in its range that the beacon is there. It does not know how many beacons or receiving devices are in the area, and it does not connect with them. An example of beacon operation is shown in Figure 3.

Beacons broadcast signals at a certain interval and within a certain transmission range. A beacon broadcasts a signal to all nearby devices that can receive the Bluetooth signal, i.e., the devices that have a Bluetooth receiver and the receiver is on. To collect the signal from the beacon, it is necessary to have a device with a BLE receiver. This can be a smartphone or a single-board computer, such as a Raspberry Pi. Applications or functions can be implemented based on the signal from the beacons. However, these applications are running on the hosting device, i.e., a smartphone or a Raspberry Pi, and not on the beacon.

Beacons are using BLE. The way the peripheral device announces its existence to the other devices is the opposite of how it is in the original Bluetooth classic. BLE enables a peripheral device to transmit an advertisement packet without being paged by the master/central device [24]. Due to this communication model, it is possible to construct energy-efficient transmitters. Moreover, when two BLE 4.0 devices are paired, they waste less battery power because the connection is dormant unless critical data are being shared. With the previous generation of Bluetooth, it was best to shut down your hardware when it was not in use. The Bluetooth SIG estimates between one and two years of battery power in some devices with Bluetooth 4.0.

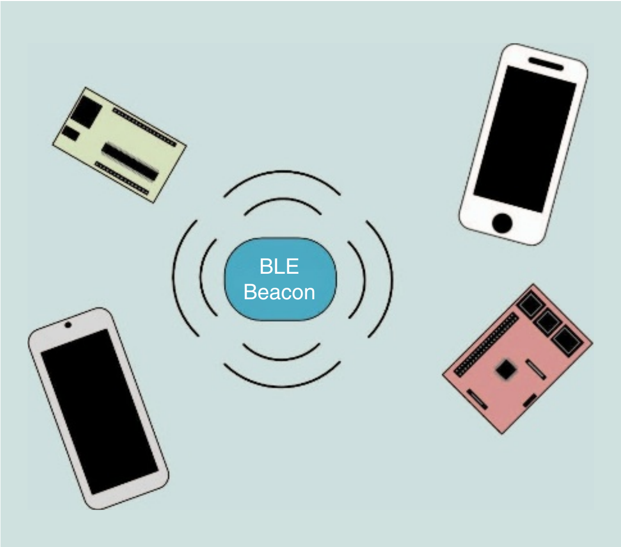


FIGURE 3. A BLE beacon broadcasting a signal to nearby devices. Each device can receive the signal and take an action in response.

Configuration parameters

BLE beacons have configuration parameters and a set of values that can determine their performance and utility for different applications. Some of these parameters are important when beacons are used in microlocation applications.

Transmission power

Transmission power is the required power to broadcast the beacon signal. As in every wireless device, transmission power directly affects the transmission range. The higher the transmission power, the longer the signal range of the beacon. This is an important tradeoff for most beacon applications. Technically, a beacon's range can reach up to 70 m. However, the battery might last for only six months. If the transmission range is constrained to 2 m, then the beacon might go up to two years without the need for battery replacement. A small transmission power can also increase the required number of beacons to cover an area, whereas a large transmission power can increase the collisions and interference. As can be inferred, an optimal transmission range can help to extend the lifetime of the beacons and minimize the battery replacement cost. At the same time, it can minimize unnecessary collisions with other beacons in the area.

Advertising interval

Advertising interval is another characteristic that affects the overall performance of beacons. It describes the time between consecutive transmissions. Applications that need to notify or detect the users that are moving in the area require a short advertising interval, and applications where the users are moving less frequently might improve their performance with a longer advertising interval. Similar to transmission power, the advertising interval affects beacon performance. The shorter the interval, the more stable the signal from the beacon. At the same time, the shorter the interval, the higher the power consumption. Once again, there is a tradeoff between beacon performance and power consumption.

BLE beacon protocols

Beacon protocols are standards of BLE communication. Each protocol describes the structure of the advertisement packet beacon's broadcast. It is necessary for the advertisement packet to have the media access control address of the beacon. There are different protocols, the most popular of which are the following.

- *iBeacon*: Apple's iBeacon was the first BLE beacon technology to come out [4]. iBeacon is a proprietary, closed standard. It broadcasts four pieces of information: 1) a universally unique identifier that identifies the beacon, 2) a major number identifying a subset of beacons within a large group, 3) a minor number identifying a specific beacon within the subset, and 4) a transmission power level in the major number's complement, indicating the signal strength 1 m from the device. This number must be calibrated for each device by the user or manufacturer. iBeacon has a simple implementation and large documentation, but it has fewer features in comparison with the following protocols. iBeacon works with iOS and Android but is native to iOS.

- *Eddystone*: Eddystone was announced from Google, and it is another protocol that defines a BLE message format for proximity beacon messages [5]. Eddystone protocol is able to transmit four different frame types: 1) a unique identifier, which is used to identify the individual beacon; 2) a uniform resource locator, which can be a website link that redirects to a website that is secured using secure sockets layer, eliminating the need for a mobile app; 3) telemetry, which includes sensor and administrative data from the beacon through telemetry, e.g., the beacon's battery level and its temperature; and 4) an encrypted identifier, which is an encrypted ephemeral identifier that changes periodically at a rate determined during the initial registration with a web service. This frame type is intended for use in security- and privacy-enhanced devices. Eddystone also works with both iOS and Android.
- *AltBeacon*: AltBeacon is an open-source beacon protocol [25] that was designed by Radius Networks. It has the same functionality as an iBeacon, but it is not company specific. This makes AltBeacon compatible with any mobile operating platform and more flexible because it has a customizable source code.
- *GeoBeacon*: GeoBeacon is another open-source beacon protocol, designed for usage in geocaching applications [26]. It has a very compact type of data storage. GeoBeacon can provide high-resolution coordinates, and it is also compatible with different mobile operating platforms.

Hardware solutions

There are a great variety of BLE beacon devices on the market. Most of them operate on batteries, such as Estimote, Kontakt, Gimbal, Glimworm, and BlueCats [27], but there are also solar-power beacons, such as the CYALKIT-E02. Each has its own unique features, such as additional sensors, battery life, reconfigurability, and dimensions, though all fundamentally work the same.

At the physical layer, BLE transmits in the 2.4-GHz industrial, scientific, and medical band with 40 channels, each 2-MHz wide. From those channels, 37 are used to exchange the data among paired devices, and three channels are designated for broadcasting advertisements. These three channels are primarily used by beacons and are chosen deliberately to minimize any collision with the Wi-Fi channels. A beacon broadcasts its advertisement packet repetitively based on the selected advertising interval while hopping over the three designated channels [28].

Beacon advantages for microlocation

Beacons have several advantages for use for microlocation.

- *Size*: Beacons are small in size and hence can be placed in almost any indoor environment with no problem. They can be placed behind the ceiling, under objects, or even on the walls.
- *Energy efficiency*: The great advantage of beacons comes from the energy efficient BLE protocol. At the same time, as the market of the beacons increases, so do the different design approaches. There are small beacons that work with one single coin cell battery, there are beacons with two AA batteries, and there are solar-powered beacons [29]. The lifetime of these beacons can be up to two years without the need for battery replacement [27].
- *Cost*: Most of the beacons in the market are cheap. Many beacons can be placed in a complex indoor environment to improve microlocation with minimum cost.
- *Interferences*: Beacons use BLE, and they will not interfere with other wireless infrastructures in the area.
- *Passive mode*: Beacons are broadcasters that do nothing else besides sending a piece of information. The logic behind each signal is done by the supporting device, such as a smartphone. Beacon signals are used by applications to trigger events and call actions, allowing the users to interact with physical things. All of the implementation is done on the device, and the beacons just broadcast the signal.
- *Platform independent*: Beacons can be used with iOS and Android devices. Each platform requires different protocols that have different packet layouts, but most platforms are able to listen to the different protocols.

Using BLE beacons for microlocation

Test case

Museums and art galleries usually provide visitors with either paper booklets or audio guides. Unfortunately, interest may vary from person to person, and each visitor's experience is also related to the available time to visit most of the exhibits. Interactive and personalized museum tours need to be developed. BLE beacons as a newly emerged technology can enhance a visitor's experience through microlocation, as shown in Figure 4.

Beacons can offer museums an opportunity to provide context to visitors through a smartphone application. Microlocation technology can make locating an exhibit much easier; at the same time, it can provide personalized suggestions to the user regarding the available exhibits. A mobile application can be developed that interacts with the available beacons.

When visitors are close to an exhibit, they can get all of the necessary information about the exhibit on their smartphone or

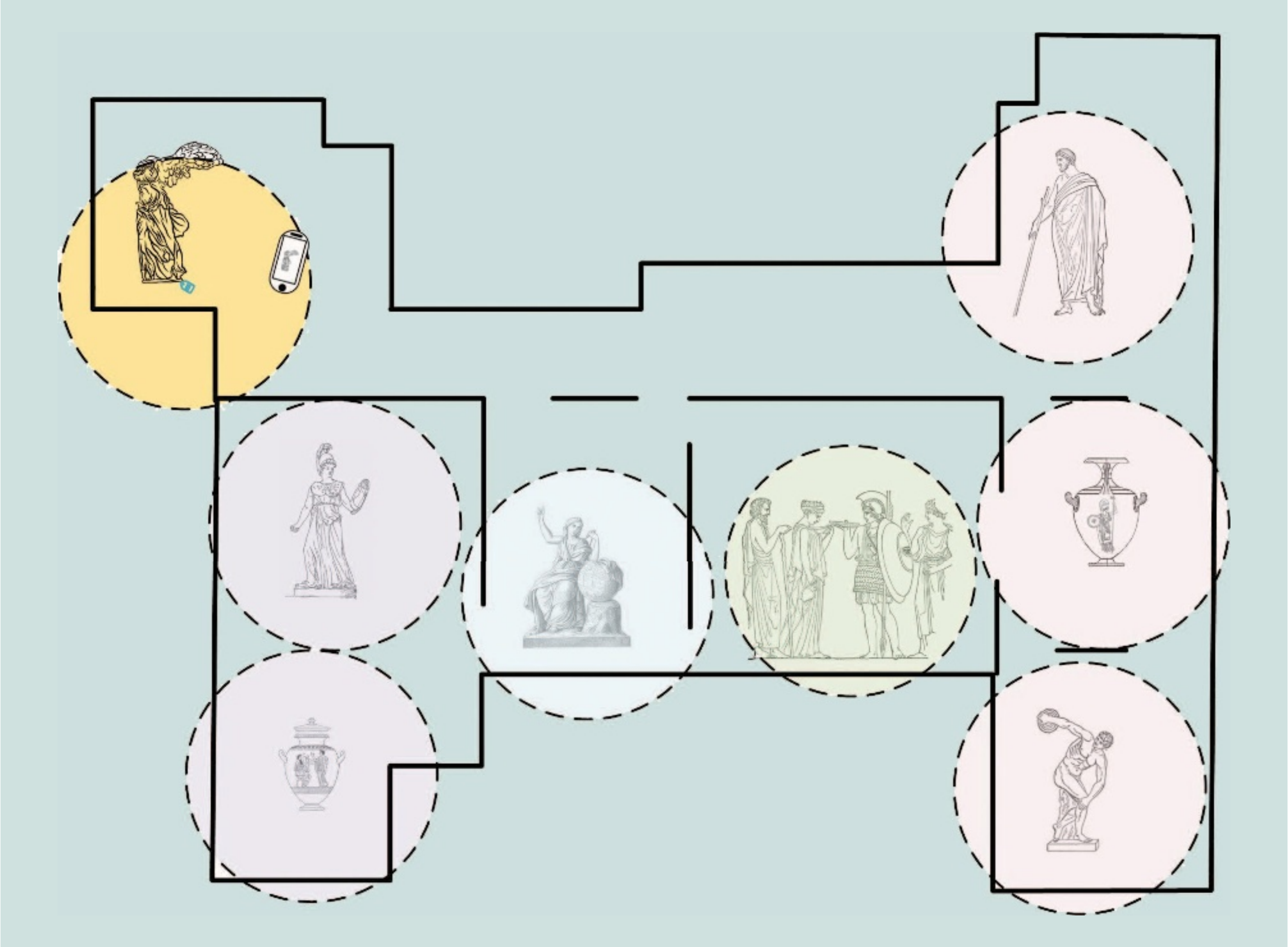


FIGURE 4. The BLE beacons used in an interactive museum scenario.

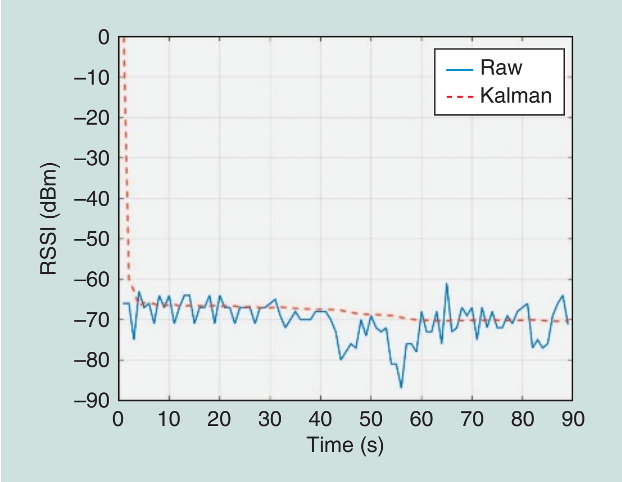


FIGURE 5. The received RSSI values 2 m from the BLE beacon.

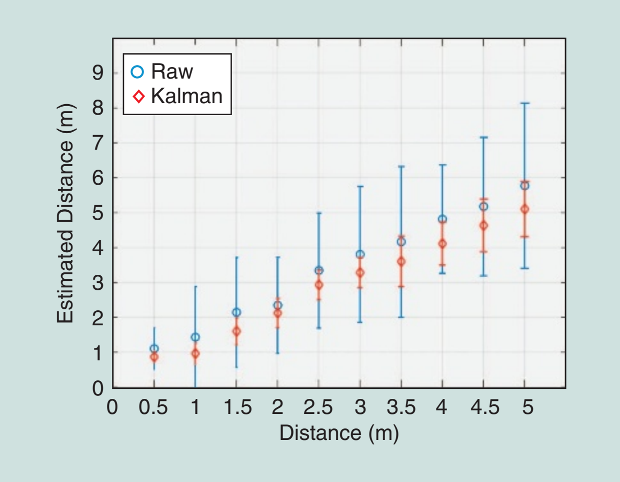


FIGURE 6. The distance estimation in ten different spots from the BLE beacon.

BLE-enabled mobile device in general. The application can also provide a recommendation to the visitor on the next exhibit he or she can visit, based on current location and interest. At the same time, the application can provide an optimal tour of the museum based on each individual's preferences. Beacons will also provide useful analytics to the museum. The number of visitors per exhibit can be collected, without violating visitor privacy. These analytics can be used to improve exhibit visibility.

The use of beacons provides several advantages for the museum and the visitors.

- *Promote exploration*: The application can encourage users to visit exhibits in different places of the museum.

Usually, visitors tend to spend most of their time in exhibits near the entrance, missing the opportunity to explore exhibits across all of the museum. Microlocation can help them identify more quickly the rooms in which they are interested.

- *Personalized tour*: When a user is interested in an exhibit, the application can provide a guided tour based on that interest. An interactive and personalized tour with exhibits from the same chronological period or within the same interest category can be provided to the user, who might miss them without the application.
- *Tour optimization*: For many visitors, the available time to spend in the museum is limited. The real-time analytics from the beacons can be used to provide an optimal route for the visitor, based on the available time for the visit.
- *Data analytics*: Beacon analytics can be used to improve the general visitor experience. There are exhibits that are missed due to their location, and there are exhibits that are overcrowded during a specific time of the day. Analytics can be used to optimize both cases and enhance the visitor experience.

Experimental results

In this section, we showcase the performance of the BLE beacons through a simple experimentation. We used BLE beacons from Gimbal Series 21 to examine the proximity estimation performance along with a smartphone, which was used to collect the signals [30]. The Kalman filter was applied on the collected data offline.

The Kalman filter estimation is shown in Figure 5. These are the collected RSSI values when the smartphone is 2 m away from the beacon. It is clear that the Kalman filter can minimize the effect of interference between the beacon and the smartphone, such as when people are moving between the two communicating devices.

To examine the performance of the Kalman filter, we placed the smartphone at ten different distances, starting from 50 cm and up to 5 m, increasing the distance 50 cm every time. In every location, we collected data on the smartphone for approximately 2 min. The average RSSI values are shown in Figure 6. When the smartphone is close to the beacon, the accuracy is high enough without filtering. As the distance increases, the accuracy without filtering decreases, and the standard deviation of the data increases as well. Interference and noise affect the data transmission; hence, as the distance between the communicating devices increases, these factors increase as well. Kalman filtering helps to keep the data close to the real value, and the standard deviation is smaller. The use of Kalman filtering helps minimize the effect of random noise and interference during the experiment.

We further examined the error between the estimated distance and the real distance and the number of occurrences of each group of errors, as shown in Figure 7. Without filtering, the error is within 3 m from the real location when distances up to 5 m are tested. In many applications that use microlocation, such as the test case, the location error should be smaller. A smaller error comes when the Kalman filter is used. The error is within 1 m from the real location, which can be acceptable for many microlocation applications.

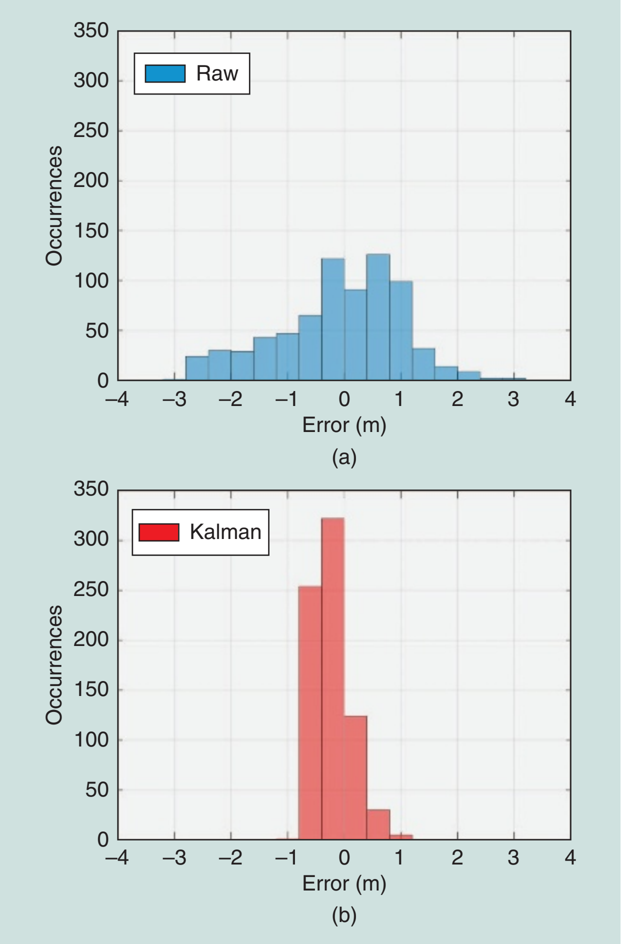


FIGURE 7. A histogram of experimental error: (a) raw data and (b) Kalman filter.

Concluding remarks

This article provides an overview of wireless technologies that can be used for microlocation in smart buildings with the use of IoT devices. BLE is among the most energy-efficient technologies. BLE beacons are small, low-cost devices that can be used for localization. Unfortunately, they are prone to interference due to their wireless nature. Signal processing techniques, such as Kalman filters, can be used to enhance their performance.

A case study of BLE beacons in an interactive museum was also discussed. According to the experimental results, signal processing techniques can enhance beacon performance and provide accurate microlocation in the era of the IoT.

Authors

Petros Spachos (petros@uoguelph.ca) received his diploma degree in electronic and computer engineering from the Technical University of Crete, Greece, in 2008 and his M.A.Sc. and Ph.D. degrees, both in electrical and computer engineering, in 2010 and 2014, respectively, from the University of Toronto, Canada. He is an assistant professor at the School of Engineering, University of Guelph, Canada. His research interests include experimental wireless networking and mobile computing with a focus on wireless sensor networks, smart cities, and the Internet of Things. He is a Senior Member of the IEEE.

Ioannis Papapanagiotou (ipapapa@ncsu.edu) received his diploma degree in electrical and computer engineering from the University of Patras, Greece, in 2006, his M.Sc. degree in computer engineering from North Carolina State University, Raleigh, in 2009, and his dual-major Ph.D. degree in computer engineering and operations research from North Carolina State University in 2012. He is an engineering manager at Netflix, Los Gatos, California, a research assistant professor at the University of New Mexico, a graduate faculty member at Purdue University, West Lafayette, Indiana, and a mentor at International Accelerator, Austin, Texas. His main focus is on distributed systems, cloud computing, and the Internet of Things. He has been awarded the NetApp faculty fellowship and established the NVIDIA CUDA Research Center at Purdue University. He has also received the IBM Ph.D. and Academy of Athens Ph.D. Fellowships for his research and the Best Paper Award at several IEEE conferences for his academic contributions. He has authored a number of research articles and patents. He is a senior member of the Association for Computing Machinery. He is a Senior Member of the IEEE.

Konstantinos N. Plataniotis (kostas@ece.utoronto.ca) received his diploma in computer engineering from the University of Patras, Greece, in 1988 and his Ph.D. in electrical engineering from the Florida Institute of Technology, Melbourne, in 1994. He is a professor and Bell Canada Chair in Multimedia with the Electrical and Computer Engineering Department at the University of Toronto, Canada. He is a Registered Professional Engineer in Ontario and fellow of the Engineering Institute of Canada. He has served as the IEEE Signal Processing Society vice president of membership (2014–2016) and the editor-in-chief of *IEEE Signal Processing Letters* (2009–2011). He was technical cochair for the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and the general cochair for the 2017 IEEE Global Conference on Signal and Information Processing. He is a cochair of the 2018 IEEE International Conference on Image Processing and ICASSP 2021. He is a Fellow of the IEEE.

References

[1] S. Sadowski and P. Spachos. (2018). RSSI-based indoor localization with the Internet of Things. [Online]. Available: https://ieee-dataport.org/documents/rssi-based-indoor-localization-internet-things

[2] F. Zafari, I. Papapanagiotou, and K. Christidis, "Microlocation for Internet-of-Things-equipped smart buildings," *IEEE Internet Things J.*, vol. 3, no. 1, pp. 96–112, Feb. 2016.

[3] D. Sikeridis, B. P. Rimal, I. Papapanagiotou, and M. Devetsikiotis, "Unsupervised crowd-assisted learning enabling location-aware facilities," *IEEE Internet Things J.*, to be published.

[4] Apple. (2014, June 2). Getting started with iBeacon. [Online]. Available: https://developer.apple.com/ibeacon/Getting-Started-with-iBeacon.pdf

[5] Google. (2017, July 5). Google Eddystone format. [Online]. Available: https://developers.google.com/beacons/eddystone

[6] J. Bordoy, A. Traub-Ens, A. Sadr, J. Wendeberg, F. Höflinger, C. Schindelhauer, and L. Reindl, "Bank of Kalman filters in closed-loop for robust localization using unsynchronized beacons," *IEEE Sensors J.*, vol. 16, no. 19, pp. 7142–7149, Oct. 2016.

[7] F. Zafari, I. Papapanagiotou, M. Devetsikiotis, and T. Hacker. (2017). An iBeacon based proximity and indoor localization system. [Online]. Available: https://arxiv.org/abs/1703.07876

[8] F. Zafari and I. Papapanagiotou, "Enhancing iBeacon based micro-location with particle filtering," in *Proc. 2015 IEEE Global Communications Conf. (GLOBECOM)*, pp. 1–7.

[9] C. Takahashi and K. Kondo, "Accuracy evaluation of an indoor positioning method using iBeacons," in *Proc. 2016 IEEE 5th Global Conf. Consumer Electronics*, pp. 1–2.

[10] Z. Chen, Q. Zhu, H. Jiang, and Y. C. Soh, "Indoor localization using smartphone sensors and iBeacons," in *Proc. 2015 IEEE 10th Conf. Industrial Electronics and Applications (ICIEA)*, pp. 1723–1728.

[11] Institute for Building Efficiency. (2018). [Online]. Available: http://www.buildingefficiencyinitiative.org/

[12] The Working Group for WLAN Standards. (2018). IEEE 802.11 wireless local area networks. [Online]. Available: http://www.ieee802.org/11/

[13] P. Baronti, P. Pillai, V. W. Chook, S. Chessa, A. Gotta, and Y. F. Hu, "Wireless sensor networks: A survey on the state of the art and the 802.15.4 and Zigbee standards," *Comput. Commun.*, vol. 30, no. 7, pp. 1655–1695, 2007.

[14] R. Want, "An introduction to RFID technology," *IEEE Pervasive Computing*, vol. 5, no. 1, pp. 25–33, Jan. 2006.

[15] Bluetooth Special Interest Group. (2018). Bluetooth 5.0 core specification. [Online]. Available: https://www.bluetooth.com/specifications/bluetooth-core-specification/bluetooth5

[16] M. Kusens, "Electronic location identification and tracking system with beacon clustering," U.S. Patent 9774991, Sept. 26, 2017.

[17] E. Khorov, A. Lyakhov, A. Krotov, and A. Guschin, "A survey on IEEE 802.11ah: An enabling networking technology for smart cities," *Comput. Commun.*, vol. 58, no. 0140-3664, pp. 53–69, 2015.

[18] Lora Alliance Technology. (2018). LoRaWAN. [Online]. Available: https://lora-alliance.org/about-lorawan

[19] B. Kennedy, G. Taylor, and P. Spachos, "BLE beacon-based patient tracking in smart care facilities," in *Proc. 2018 IEEE Int. Conf. Pervasive Computing and Communications Workshops (PerCom Workshops)*.

[20] M. Ayyash, H. Elgala, A. Khreishah, V. Jungnickel, T. Little, S. Shao, M. Rahaim, D. Schulz, et al., "Coexistence of WiFi and LiFi toward 5G: Concepts, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 64–71, Feb. 2016.

[21] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, 2002.

[22] P. M. Djuric, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Miguez, "Particle filtering," *IEEE Signal Process. Mag.*, vol. 20, no. 5, pp. 19–38, 2003.

[23] I. Guvenc, C. Abdallah, R. Jordan, and O. Dedoglu, "Enhancements to RSS-based indoor tracking systems using Kalman filters," in *Proc. Int. Signal Processing Conf.*, 2003.

[24] F. M. P. Kriz and T. Kozel, "Improving indoor localization using Bluetooth low energy beacons," *Mobile Inform. Syst.*, vol. 2016, pp. 1–7, Mar. 2016.

[25] AltBeacon. (2016, June 7). Specifications on AltBeacon. [Online]. Available: https://github.com/AltBeacon/spec

[26] GeoBeacon. (2018). Tecno world. [Online]. Available: https://github.com/Tecno-World/GeoBeacon

[27] Aislelabs. (2015). The hitchhiker's guide to iBeacon hardware: A comprehensive report. [Online]. Available: http://www.aislelabs.com/reports/beacon-guide/

[28] E. Vlugt. (2013). Bluetooth low energy, beacons and retail. [Online]. Available: http://global.verifone.com/media/3603729/bluetooth-low-energy-beacons-retail-wp.pdf

[29] P. Spachos and A. Mackey, "Energy efficiency and accuracy of solar powered BLE beacons," *Comput. Commun.*, vol. 119, pp. 94–100, Apr. 2018.

[30] A. Mackey and P. Spachos, "Performance evaluation of beacons for indoor localization in smart buildings," in *Proc. 2017 IEEE Global Conf. Signal and Information Processing (GlobalSIP)*, pp. 823–827.

Moe Z. Win, Florian Meyer, Zhenyu Liu, Wenhan Dai,
Stefania Bartoletti, and Andrea Conti

Efficient Multisensor Localization for the Internet of Things

Exploring a new class of scalable localization algorithms

INTERNET OF THINGS—ISTOCKPHOTO.COM/IAREMENKO
CIRCUITS—IMAGE LICENSED BY INGRAM PUBLISHING

In the era of the Internet of Things (IoT), efficient localization is essential for emerging mass-market services and applications. IoT devices are heterogeneous in signaling, sensing, and mobility, and their resources for computation and communication are typically limited. Therefore, to enable location awareness in large-scale IoT networks, there is a need for efficient, scalable, and distributed multisensor fusion algorithms. This article presents a framework for designing network localization and navigation (NLN) for the IoT. Multisensor localization and operation algorithms developed within NLN can exploit spatiotemporal cooperation, are suitable for arbitrary, large-network sizes, and only rely on an information exchange among neighboring devices. The advantages of NLN are evaluated in a large-scale IoT network with 500 agents. In particular, because of multisensor fusion and cooperation, the presented network localization and operation algorithms can provide attractive localization performance and reduce communication overhead and energy consumption.

IoT location awareness

Location awareness [1]–[6] is a cornerstone of the IoT and fosters a wide range of emerging applications, such as crowdsensing [7], big data analysis [8], environmental monitoring [9], and autonomous driving [10]. The position information of IoT devices can contribute to connecting and exchanging data more efficiently, preserving communication security, and allowing autonomous motion. The increasing number and different types of IoT devices generate scenarios in which heterogeneous data are collected distributedly using different sensing technologies. Compared to conventional wireless localization networks that typically consist of a limited number of homogeneous nodes, the scale and heterogeneity of an IoT network imposes new challenges that need to be addressed. Specifically, IoT localization and navigation calls for a new class of algorithms tailored to IoT networks.

In IoT networks, the sensing capabilities of the devices can vary significantly, providing different kinds of measurements carrying positional information such as range, angle of arrival,

Digital Object Identifier 10.1109/MSP.2018.2845907
Date of publication: 28 August 2018

channel state information, or inertial. Additionally, depending on the specific sensing technology used by each device, communication ranges and measurement accuracies are different. Since IoT devices are typically equipped only with inexpensive sensors having limited capabilities, high-accuracy localization and navigation usually requires multisensor fusion and device cooperation. However, state-of-the-art multisensor fusion algorithms based on sequential Bayesian estimation (SBE) [11]–[13] are often impractical for IoT applications due to their decentralized network topology and the limited processing units of IoT devices. Moreover, the high number of devices necessitates network operation strategies that provide interdevice cooperation for an efficient use of the limited battery power and spectral resources. For these reasons, the major difficulties for efficient multisensor localization and navigation in the IoT lie in fusing data and measurements collected from heterogeneous sensors with low computation and communication capabilities and in designing network operation strategies that can efficiently allocate resources in scenarios with insufficient infrastructure and limited battery power. Addressing these difficulties can overcome key issues in the current IoT networks, including the heterogeneity of sensing technologies and the limited capability of devices in terms of computation, communication, and battery energy.

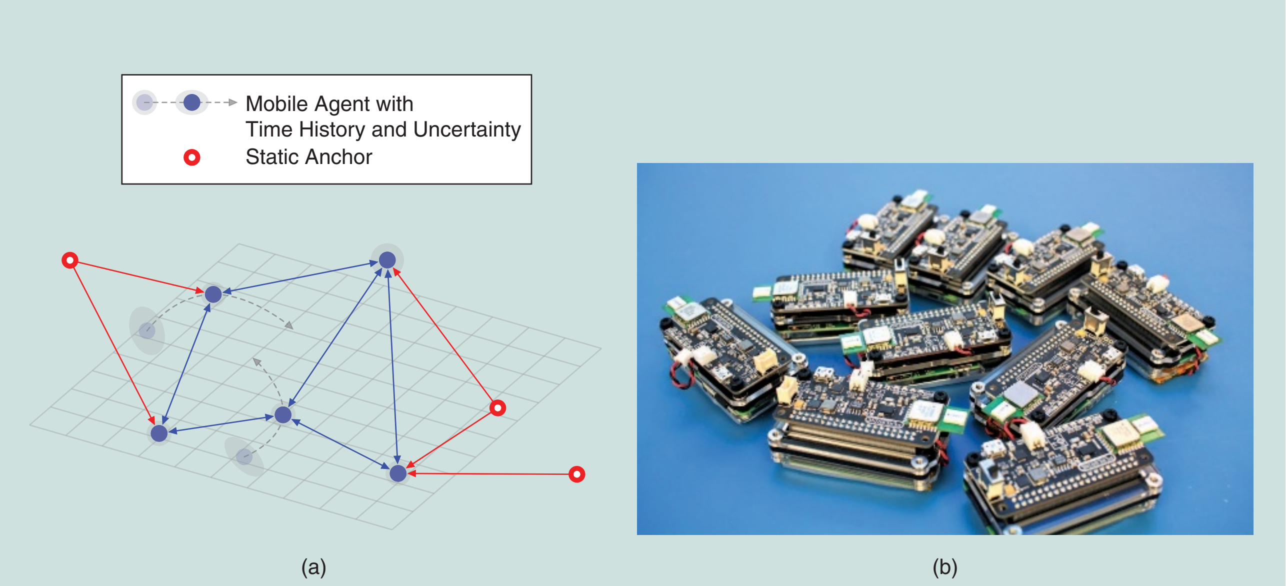
The recently introduced paradigm of NLN [1] has important characteristics that are favorable for multisensor localization and navigation in IoT networks. In particular, it can provide technology-agnostic and low-complexity algorithms for heterogeneous multisensor fusion [14] and scalable network operation [15], which typically do not require much communication and computation overhead. An NLN scenario involving five devices and three anchors is shown in Figure 1(a). Figure 1(b) shows devices of Peregrine, a system developed for a three-dimensional (3-D) NLN.

This article provides an overview of how IoT location awareness can be enabled by the NLN paradigm.

- We present a framework for developing scalable and distributed inference algorithms for localization in IoT networks.
- We devise centralized and distributed network operation strategies that can increase battery lifetime and localization accuracy.
- We demonstrate that multisensor fusion and cooperation among devices can dramatically increase localization performance in a large-scale scenario with hundreds of mobile agents.
- We quantify how network operation algorithms can reduce the communication overhead and energy consumption of localization networks.

Notation

Random variables (RVs) are displayed in sans serif, upright fonts; their realizations in serif, italic fonts. Vectors and matrices are denoted by bold lowercase and uppercase letters, respectively. Sets are denoted by calligraphic font. For example, an RV and its realization are denoted by x and x, respectively; a random vector and its realization are denoted by $\boldsymbol{\mathsf{x}}$ and $\boldsymbol{x}$, respectively; a set is denoted by $\mathcal{X}$. The identity matrix is denoted by $\boldsymbol{I}$. For the probability distribution function (PDF) of the random vector $\boldsymbol{\mathsf{x}}$, at $\boldsymbol{x}$, the short notation $f(\boldsymbol{x}) = f_{\boldsymbol{\mathsf{x}}}(\boldsymbol{x})$ is used. Furthermore, $\boldsymbol{x} = [\boldsymbol{x}_i]_{i \in \mathcal{I}}$ denotes vector that is obtained by arranging all the subvectors $\boldsymbol{x}_i, i \in \mathcal{I}$ in an arbitrary but known order into a column vector. Finally, the notations of important quantities that are used throughout the article are summarized in Table 1.



(a)

(b)

FIGURE 1. (a) A graphical depiction of an NLN scenario involving five devices and three anchors. (b) The devices used in the Peregrine, a system for a 3-D NLN [16].

Table 1. Notations of important quantities.

Notation	Definition	Notation	Definition
$\mathcal{N}_{\mathrm{a}}$	The index set of mobile agents	$\mathcal{N}_{\mathrm{b}}$	The index set of anchors
$\mathbf{x}_i^{(n)}$	The positional state of the *i*th node at time *n*	$\mathbf{p}_i^{(n)}$	The position of the *i*th node at time *n*
$\mathbf{z}_{ij}^{(n)}$	An internode measurement between *i*th agent and *j*th node at time *n*	$\mathbf{z}_i^{(n)}$	All the internode measurements of the *i*th agent at time *n*
$\mathbf{x}_i^{(0:n)}$	All the positional states of the *i*th node up to time *n*	$\mathbf{z}_i^{(1:n)}$	All the measurements of the *i*th agent up to time *n*
$\alpha_f(\boldsymbol{x}^{(n)})$	The message passed from variable node $\boldsymbol{x}$ to factor node *f*	$\beta_f(\boldsymbol{x}^{(n)})$	The message passed from factor node *f* to variable node $\boldsymbol{x}$
$\boldsymbol{\mu}_{\mathrm{p}}^{(n)}$	The predicted mean vector	$\Sigma_{\mathrm{p}}^{(n)}$	The predicted covariance matrix
$\boldsymbol{\mu}^{(n)}$	The posterior mean vector	$\Sigma^{(n)}$	The posterior covariance matrix
$\bar{\mathbf{x}}_i^{(n)}$	The augmented state vector	$\bar{\mathbf{z}}_i^{(n)}$	The augmented measurement vector
$\boldsymbol{Q}^{(n)}$	The localization error matrix	$\boldsymbol{J}^{(n)}$	The Fisher information matrix
$\mathcal{P}_{\mathrm{NA}}^{(n)}$	The optimization problem for node activation	$\mathcal{P}_{\mathrm{NP}}^{(n)}$	The optimization problem for node prioritization
$\zeta_i^{(n)}$	The channel access probability of agent *i*	$y_{ij}^{(n)}$	The amount of resources allocated to the measurement link pair (*i*, *j*)
$\chi_i^{(n)}$	The potential error reduction of agent *i* related to internode measurements	$\xi_{ij}^{(n)}$	The channel quality between nodes *i* and *j*

Single-node localization for IoT

This section revises localization and navigation algorithms for single-node scenarios. First, consider a network of IoT devices that consists of a mobile agent (with index set $\mathcal{N}_a = \{1\}$) and of N_b anchors at known positions (with index set $\mathcal{N}_{\mathrm{b}} = \{2, 3, ..., N_{\mathrm{b}} + 1\}$). The agents are localized based on heterogeneous sensor measurements by using the anchors as reference points. Measurements for localization are made at discrete time steps indexed by $n = 1, 2, ..., N$. Let $\mathbf{x}_1^{(n)} \in \mathbb{R}^D$ be the unknown positional state of the agent at time n, which includes the position $\mathbf{p}_1^{(n)}$ and other mobility parameters such as velocity, acceleration, orientation, and angular velocity. All measurements made at time n are summarized in the vector $z_1^{(n)}$, which is the concatenation of all internode measurements $z_{1j}^{(n)}$ with anchors $j \in \mathcal{N}_{\mathrm{b}}$. The localization process is essentially the calculation of an estimate $\hat{x}_1^{(n)}$ of $\mathbf{x}_1^{(n)}$ from all available measurements up to time n (denoted as $z_1^{(1:n)} \triangleq [z_1^{(1)\mathrm{T}}, z_1^{(2)\mathrm{T}}, \ldots, z_1^{(n)\mathrm{T}}]^{\mathrm{T}}$).

The relationship of the current state vector with the previous state vector can be described by the state-evolution model

$$\mathbf{x}_1^{(n)} = \boldsymbol{a}(\mathbf{x}_1^{(n-1)}, \mathbf{c}_1^{(n)}; \boldsymbol{u}_1^{(n)}), \tag{1}$$

where $\mathbf{c}_1^{(n)}$ is the state-evolution noise vector that is assumed independent across time n and $\boldsymbol{u}_1^{(n)}$ is a known input [17] that controls the motion of the agent. Note that the PDF $f(\boldsymbol{c}_1^{(n)})$ can be different for distinct time steps n. From the state-evolution model (1) one can directly obtain the state-evolution function $f\left(\boldsymbol{x}_1^{(n)} \middle| \boldsymbol{x}_1^{(n-1)}; \boldsymbol{u}_1^{(n)}\right)$. Note that (1) implies a Markov property, i.e., given $\boldsymbol{x}_1^{(n-1)}, \mathbf{x}_1^{(n)}$ is statistically independent of previous $\mathbf{x}_1^{(0)}, \mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_1^{(n-2)}$ and future $\mathbf{x}_1^{(n+1)}, \mathbf{x}_1^{(n+2)}, \ldots$ states. The joint prior PDF $f(\boldsymbol{x}_1^{(0)})$ at time $n = 0$ is known. The joint prior information for all times $0, 1, \ldots, n$, i.e., all available information before any measurement is performed, can now be expressed as

$$f(\boldsymbol{x}_1^{(0:n)}; \boldsymbol{u}_1^{(1:n)}) = f(\boldsymbol{x}_1^{(0)}) \prod_{k=1}^{n} f\left(\boldsymbol{x}_1^{(k)} \middle| \boldsymbol{x}_1^{(k-1)}; \boldsymbol{u}_1^{(k)}\right). \tag{2}$$

The relationship of the current measurements with the current state vector is described by the measurement model

$$\mathbf{z}_1^{(n)} = \boldsymbol{h}(\mathbf{x}_1^{(n)}, \mathbf{v}_1^{(n)}), \tag{3}$$

where $\mathbf{v}_1^{(n)}$ is the measurement noise, which is assumed independent across times n. Note that the PDF $f(\boldsymbol{v}_1^{(n)})$ can be different for distinct time steps n. From the measurement model (3) one can directly obtain the likelihood function $f\left(z_1^{(n)} \middle| \boldsymbol{x}_1^{(n)}\right)$. Note that (3) implies that given $\boldsymbol{x}_1^{(n)}, \mathbf{z}_1^{(n)}$ is statistically independent of previous $\mathbf{x}_1^{(0)}, \mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_1^{(n-1)}$ and of future $\mathbf{x}_1^{(n+1)}, \mathbf{x}_1^{(n+2)}, \ldots$ states, as well as of previous $\mathbf{z}_1^{(0)}, \mathbf{z}_1^{(1)}, \ldots, \mathbf{z}_1^{(n-1)}$ and future $\mathbf{z}_1^{(n+1)}, \mathbf{z}_1^{(n+2)}, \ldots$ measurements. Therefore, the likelihood function for all times $1, 2, \ldots, n$ (i.e., all available information related to the performed measurements) can be expressed as

$$f\left(z_1^{(1:n)} \middle| \boldsymbol{x}_1^{(1:n)}\right) = \prod_{k=1}^{n} f\left(z_1^{(k)} \middle| \boldsymbol{x}_1^{(k)}\right). \tag{4}$$

By using Bayes' rules, (2) and (4), the joint posterior PDF of $\mathbf{x}_1^{(0:n)}$ given $z_1^{(1:n)}$ for $n > 0$ results in

$$\begin{aligned} & f\left(\boldsymbol{x}_1^{(0:n)} \middle| z_1^{(1:n)}; \boldsymbol{u}_1^{(1:n)}\right) \\ & \propto f\left(z_1^{(1:n)} \middle| \boldsymbol{x}_1^{(1:n)}\right) f(\boldsymbol{x}_1^{(0:n)}; \boldsymbol{u}_1^{(1:n)}) \\ & = f(\boldsymbol{x}_1^{(0)}) \prod_{k=1}^{n} f\left(\boldsymbol{x}_1^{(k)} \middle| \boldsymbol{x}_1^{(k-1)}; \boldsymbol{u}_1^{(k)}\right) f\left(z_1^{(k)} \middle| \boldsymbol{x}_1^{(k)}\right). \end{aligned} \tag{5}$$

The factor graph [18] representing this joint posterior for SBE is shown in Figure 2. For simplicity in notation, the index of the agent is dropped in the following, e.g., $\mathbf{x}_1^{(n)}$ is replaced by $\mathbf{x}^{(n)}$.

A temporal fusion based on SBE

Temporal multisensor fusion in a Bayesian setting is accomplished by determining an estimate of $\mathbf{x}^{(n)}$ from the marginal posterior PDF $f\left(\boldsymbol{x}^{(n)}\middle|\boldsymbol{z}^{(1:n)}\right)$. For example, the minimum mean-square-error (MMSE) estimate is given by [19]

$$\hat{\boldsymbol{x}}^{(n)}_{\mathrm{MMSE}} \triangleq \int \boldsymbol{x}^{(n)} f\left(\boldsymbol{x}^{(n)}\middle|\boldsymbol{z}^{(1:n)};\boldsymbol{u}^{(1:n)}\right)\mathrm{d}\boldsymbol{x}^{(n)}. \tag{6}$$

The marginal posterior PDF $f\left(\boldsymbol{x}^{(n)}\middle|\boldsymbol{z}^{(1:n)};u^{(1:n)}\right)$ in (6) can be obtained from the joint posterior PDF $f\left(\boldsymbol{x}^{(1:n)}\middle|\boldsymbol{z}^{(1:n)};\boldsymbol{u}^{(1:n)}\right)$ in (5) by marginalization. However, direct marginalization of $f\left(\boldsymbol{x}^{(1:n)}\middle|\boldsymbol{z}^{(1:n)};\boldsymbol{u}^{(1:n)}\right)$ is unfeasible in general because it relies on integration over a state space whose dimension grows with the time n.

This problem known as the *curse of dimensionality* [20], can be addressed by SBE [12] if the joint posterior PDF $f\left(\boldsymbol{x}^{(1:n)}\middle|\boldsymbol{z}^{(1:n)};\boldsymbol{u}^{(1:n)}\right)$ has a structure like (5). The exact calculation of $f\left(\boldsymbol{x}^{(n)}\middle|\boldsymbol{z}^{(1:n)};\boldsymbol{u}^{(1:n)}\right)$ is then possible sequentially; at each time n, SBE consists of the prediction step

$$\begin{aligned} &f\left(\boldsymbol{x}^{(n)}\middle|\boldsymbol{z}^{(1:n-1)};\boldsymbol{u}^{(1:n)}\right)\\ &= \int f\left(\boldsymbol{x}^{(n)}\middle|\boldsymbol{x}^{(n-1)};\boldsymbol{u}^{(n)}\right) f\left(\boldsymbol{x}^{(n-1)}\middle|\boldsymbol{z}^{(1:n-1)};\ \boldsymbol{u}^{(1:n-1)}\right)\mathrm{d}\boldsymbol{x}^{(n-1)}, \end{aligned} \tag{7}$$

which is followed by the update step

$$f\left(\boldsymbol{x}^{(n)}\middle|\boldsymbol{z}^{(1:n)};\boldsymbol{u}^{(1:n)}\right) \propto f\left(\boldsymbol{z}^{(n)}\middle|\boldsymbol{x}^{(n)}\right) f\left(\boldsymbol{x}^{(n)}\middle|\boldsymbol{z}^{(1:n-1)};\boldsymbol{u}^{(1:n)}\right). \tag{8}$$

Contrary to direct marginalization in which integration is performed over an nD-dimensional state space, SBE involves only operations in D-dimensional state spaces that are performed n times. As a consequence, the complexity related to calculating $f\left(\boldsymbol{x}^{(n)}\middle|\boldsymbol{z}^{(1:n-1)};\boldsymbol{u}^{(n)}\right)$ scales only linearly with the number of time steps n. Note that the information acquired by all sensors up to time n, is represented by the low-dimensional predicted posterior PDF $f\left(\boldsymbol{x}^{(n)}\middle|\boldsymbol{z}^{(1:n-1)};\boldsymbol{u}^{(n)}\right)$ and temporal fusion is directly performed in the update step according to (8).

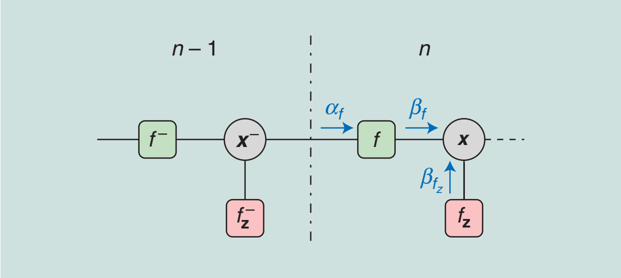


FIGURE 2. A factor graph for single-node localization representing the factorization in (5). Nodes in green represent factors related to the state-evolution function, nodes in red represent factors related to the likelihood function, while messages related to the SPA are in blue. The following short notations are used: $\boldsymbol{x}^- \triangleq \boldsymbol{x}_1^{(n-1)}$, $\boldsymbol{x} \triangleq \boldsymbol{x}_1^{(n)}$, $f^- \triangleq f\left(\boldsymbol{x}_1^{(n-1)} \mid \boldsymbol{x}_1^{(n-2)};\boldsymbol{u}_1^{(n-1)}\right)$, $f \triangleq f\left(\boldsymbol{x}_1^{(n)} \mid \boldsymbol{x}_1^{(n-1)};\boldsymbol{u}_1^{(n)}\right)$, $f_z^- \triangleq f\left(\boldsymbol{z}_1^{(n-1)} \mid \boldsymbol{x}_1^{(n-1)}\right)$, $f_z \triangleq f\left(\boldsymbol{z}_1^{(n)} \mid \boldsymbol{x}_1^{(n)}\right)$, $\alpha_f \triangleq \alpha_f\left(\boldsymbol{x}_1^{(n-1)}\right)$, $\beta_f \triangleq \beta_f\left(\boldsymbol{x}_1^{(n)}\right)$, and $\beta_{f_z} \triangleq \beta_{f_z}\left(\boldsymbol{x}_1^{(n)}\right)$.

Message-passing interpretation of SBE

For an arbitrary estimation problem, the sum-product algorithm (SPA) [18] can calculate exact or approximate marginal posterior PDFs in an efficient manner. In particular, the SPA avoids the curse of dimensionality inherent to direct marginalization. Therefore, SPA-based solutions are attractive for high-dimensional inference problems. The SPA is a message-passing algorithm since its basic operations can be interpreted as an exchange of statistical information on adjacent nodes of a factor graph, i.e., as messages passed along the edges of the graph.

If the factor graph is tree structured, such as the one shown in Figure 2, message updates are performed only once for each node in the graph. The message-passing procedure begins at the variable and factor nodes with only one edge (which passes a constant message and the corresponding factor, respectively) and continues with those nodes where all incoming messages are computed already. According to the SPA message-passing rules, in a factor graph as shown in Figure 2, the message passed from factor node f to variable node $\boldsymbol{x}$ is obtained as [18]

$$\beta_f\left(\boldsymbol{x}^{(n)}\right) = \int f\left(\boldsymbol{x}^{(n)}\middle|\boldsymbol{x}^{(n-1)};\boldsymbol{u}^{(n)}\right)\alpha_f\left(\boldsymbol{x}^{(n-1)}\right)\mathrm{d}\boldsymbol{x}^{(n-1)}, \tag{9}$$

where $\alpha_f\left(\boldsymbol{x}^{(n-1)}\right)$ is the message passed from variable node $\boldsymbol{x}^-$ to factor node f. Furthermore, the message passed from f_z to $\boldsymbol{x}$ is given by $\beta_{f_z}\left(\boldsymbol{x}^{(n)}\right) = f\left(\boldsymbol{z}^{(n)}\middle|\boldsymbol{x}^{(n)}\right)$. After these two messages are calculated, the belief for $\boldsymbol{x}$ is finally obtained as

$$b\left(\boldsymbol{x}^{(n)}\right) \propto \beta_f\left(\boldsymbol{x}^{(n)}\right)\beta_{f^z}\left(\boldsymbol{x}^{(n)}\right). \tag{10}$$

For $\alpha_f\left(\boldsymbol{x}^{(n-1)}\right) = b\left(\boldsymbol{x}^{(n-1)}\right)$, it can be seen that $b\left(\boldsymbol{x}^{(n)}\right) = f\left(\boldsymbol{x}^{(n)}\middle|\boldsymbol{z}^{(1:n)};\boldsymbol{u}^{(1:n)}\right)$ as provided by SBE. Thus, SBE based on prediction and update steps, respectively (7) and (8), is equivalent to calculating the belief $b\left(\boldsymbol{x}^{(n)}\right)$ by running the SPA on the factor graph in Figure 2.

Node localization and navigation algorithms

A large variety of filtering algorithms suitable for node localization and navigation are based on SBE according to (7) and (8). Here, we focus on two widely adopted techniques: Kalman filtering and particle filtering.

The Kalman filter

Consider the case where the state-evolution model and the measurement model are linear, i.e., (1) and (3) can be expressed as

$$\mathbf{x}^{(n)} = \boldsymbol{A}\mathbf{x}^{(n-1)} + \boldsymbol{B}\boldsymbol{u}^{(n)} + \mathbf{c}^{(n)} \tag{10a}$$

$$\mathbf{z}^{(n)} = \boldsymbol{H}\mathbf{x}^{(n)} + \mathbf{v}^{(n)}, \tag{10b}$$

where the matrices $\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{H}$ are assumed known. Furthermore, the noise $\mathbf{c}^{(n)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{c}}^{(n)})$ and $\mathbf{v}^{(n)} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{v}}^{(n)})$ is Gaussian distributed with noise covariance matrices $\Sigma_{\mathbf{c}}^{(n)}$ and $\Sigma_{\mathbf{v}}^{(n)}$. In this case, closed-form solutions for the prediction (7) and update step (8) of SBE can be obtained. These closed-from expressions are used within the Kalman filter (KF) [19] that represents posterior PDFs $f\left(\boldsymbol{x}^{(n)}\middle|z^{(1:n)};\boldsymbol{u}^{(1:n)}\right)$ by second-order statistics, i.e., by means $\boldsymbol{\mu}^{(n)}$ and covariance matrices $\Sigma^{(n)}$. If the prior $f(\boldsymbol{x}^{(0)})$ is also Gaussian, the PDFs $f\left(\boldsymbol{x}^{(n)}\middle|z^{(1:n-1)};\boldsymbol{u}^{(1:n)}\right)$ and $f\left(\boldsymbol{x}^{(n)}\middle|z^{(1:n)};\boldsymbol{u}^{(1:n)}\right)$ are Gaussian as well for arbitrary n. In that case, the KF can provide the optimum solution and the exact MMSE estimator $\hat{\boldsymbol{x}}_{\mathrm{MMSE}}^{(n)}$ in (6) is given by $\boldsymbol{\mu}^{(n)}$. The KF consists of two steps: In the prediction step of the KF, the predicted mean $\boldsymbol{\mu}_{\mathrm{p}}^{(n)}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathrm{p}}^{(n)}$ that fully characterize $f\left(\boldsymbol{x}^{(n)}\middle|z^{(1:n-1)};\boldsymbol{u}^{(1:n)}\right)$ are calculated based on (10a). In the update step of the KF, first the mean $\boldsymbol{\mu}_{\mathbf{z}}^{(n)}$, the covariance matrix $\Sigma_{\mathbf{z}}^{(n)}$, and the cross-covariance matrix $\Sigma_{\mathbf{xz}}^{(n)}$ are calculated based on (10b), then the posterior mean $\boldsymbol{\mu}^{(n)}$ and posterior covariance matrix $\boldsymbol{\Sigma}^{(n)}$ are obtained using the Kalman update equations [19]. For nonlinear non-Gaussian models inherent to multisensor localization, computationally feasible approximate algorithms include variants of the KF, such as the extended KF (EKF) [19] and the unscented KF (UKF) [11].

The EKF and the UKF are versions of the KF that are suitable for nonlinear state-evolution and measurement models. If $\boldsymbol{a}(\mathbf{x}^{(n-1)}, \mathbf{c}^{(n)}; \mathbf{u}^{(n)})$ in (1) and $\boldsymbol{h}(\mathbf{x}^{(n)}, \mathbf{v}^{(n)})$ in (3) are nonlinear functions, the covariance matrices $\boldsymbol{\Sigma}_{\mathrm{p}}^{(n)}$ as well as $\boldsymbol{\Sigma}_{\mathbf{z}}^{(n)}$ and $\boldsymbol{\Sigma}_{\mathbf{xz}}^{(n)}$ cannot be calculated directly. The EKF and the UKF are still based on the Kalman update equations but perform different approximations to obtain these matrices.

In the EKF, a multivariate Taylor series expansion of (1) and (3) is used to linearize them around $[\boldsymbol{\mu}^{(n-1)T},\mathbf{0}^{T}]^{T}$ and $[\boldsymbol{\mu}_{\mathrm{p}}^{(n)T},\mathbf{0}^{T}]^{T}$, respectively [19]. In this way, an approximation of the matrices $\boldsymbol{\Sigma}_{\mathrm{p}}^{(n)}$, $\boldsymbol{\Sigma}_{\mathbf{z}}^{(n)}$, and $\boldsymbol{\Sigma}_{\mathbf{xz}}^{(n)}$ is obtained. While the EKF is widely adopted, it is accurate only if the system model is moderately nonlinear. Furthermore, the EKF is challenging to implement and difficult to tune. The UKF is a widely adopted solution for applications in which the EKF is not accurate or (1) and (3) are not differentiable. The UKF performs approximate inference by using a minimal set of deterministically chosen samples referred to as *sigma points* (*SPs*) [11]. The nonlinear model (1) and (3) is evaluated at the SPs and from the resulting new SPs, approximate second-order statistics $\boldsymbol{\mu}_{\mathrm{p}}^{(n)}, \boldsymbol{\Sigma}_{\mathrm{p}}^{(n)}$ as well as $\boldsymbol{\mu}_{\mathbf{z}}^{(n)}, \boldsymbol{\Sigma}_{\mathbf{z}}^{(n)}$, and $\boldsymbol{\Sigma}_{\mathbf{xz}}^{(n)}$ are calculated [11]. The UKF can often provide approximations of $\boldsymbol{\mu}^{(n)}$ and $\boldsymbol{\Sigma}^{(n)}$ that are more accurate compared to those provided by the EKF at a comparable computational complexity.

The particle filter

The particle filter (PF) is an attractive alternative to the EKF and the UKF for applications in which a representation of $f\left(\boldsymbol{x}^{(n)}\middle|z^{(1:n)};\boldsymbol{u}^{(1:n)}\right)$ using second-order statistics is not accurate. This might be the case if the state-evolution and/or measurement model are highly nonlinear and $f\left(\boldsymbol{x}^{(n)}\middle|z^{(1:n)};\boldsymbol{u}^{(1:n)}\right)$ is multimodal. The key idea of PFs is to represent the posterior distribution by a set of samples (particles) with associated weights, i.e.,

$$\tilde{f}\left(\boldsymbol{x}^{(n)}\middle|z^{(1:n)};\boldsymbol{u}^{(1:n)}\right) \approx \sum_{l=1}^{n_{\mathrm{p}}} w_l^{(n)} \delta(\boldsymbol{x}^{(n)} - \boldsymbol{x}_l^{(n)}), \qquad (11)$$

where n_{p} is the number of particles, $\delta(\cdot)$ is the Dirac delta function, $w_l^{(n)} \geqslant 0$ is the weight of the lth particle $\boldsymbol{x}_l^{(n)}$ at time index n, and $\Sigma_{l=1}^{n_{\mathrm{p}}} w_l^{(n)} = 1$. Note that the number of randomly sampled particles n_{p} is typically significantly larger compared to the number of deterministically calculated SPs n_{s} used in the UKF.

An approximation of the MMSE estimate in (6) is given by the mean of $\tilde{f}\left(\boldsymbol{x}^{(n)}\middle|z^{(1:n)};\boldsymbol{u}^{(1:n)}\right)$ in (11), which is equal to the mean of the weighted particles, i.e.,

$$\hat{\boldsymbol{x}}^{(n)} = \int \boldsymbol{x}^{(n)} \tilde{f}\left(\boldsymbol{x}^{(n)}\middle|z^{(1:n)};\boldsymbol{u}^{(1:n)}\right) \mathrm{d}\boldsymbol{x}^{(n)} = \sum_{l=1}^{n_{\mathrm{p}}} w_l^{(n)} \boldsymbol{x}_l^{(n)}. \qquad (12)$$

A large variety of particle-filtering algorithms have been introduced. In what follows, we review the prominent sequential importance resampling filter [12], which consists of three steps referred to as *sampling*, *weight update*, and *resampling*.

The sampling step corresponds to the prediction step of SBE in (7). For each particle $\boldsymbol{x}_l^{(n-1)}$, a new particle $\boldsymbol{x}_l^{(n)}$ is drawn from the state-evolution PDF $f\left(\boldsymbol{x}^{(n)}\middle|\boldsymbol{x}^{(n-1)}; \boldsymbol{u}^{(n)}\right)$ evaluated at $\boldsymbol{x}_l^{(n-1)}$. The weight update step corresponds to the update step of SBE in (8). For each particle $\boldsymbol{x}_l^{(n)}$ the updated weight $w_l^{(n)}$ is obtained as $w_l^{(n)} = f\left(z^{(n)}\middle|\boldsymbol{x}_l^{(n)}\right) \big/ \Sigma_{\ell=1}^{n_{\mathrm{p}}} f\left(z^{(n)}\middle|\boldsymbol{x}_\ell^{(n)}\right)$. Then, particle-based state estimation is performed as in (12). The resampling step is a step that is performed to avoid degeneracy of particles. It is typically executed only if an indicator called the *effective sample size* is smaller than a threshold. In the resampling step, n_{p} resampled particles are obtained by sampling from $\tilde{f}\left(\boldsymbol{x}^{(n)}\middle|z^{(1:n)}; \boldsymbol{u}^{(1:n)}\right)$ in (11) and setting the weight of the resampled particles to $1/n_{\mathrm{p}}$, with resampled particles used at time $n+1$.

Remark 1

Most PFs are optimum in the sense that for $n_{\mathrm{p}} \rightarrow \infty$ the estimate $\hat{\boldsymbol{x}}^{(n)}$ in (12) converges to the true MMSE estimate $\hat{\boldsymbol{x}}_{\mathrm{MMSE}}^{(n)}$ in (6). Contrary to EKF and the UKF, PFs are also suitable for highly nonlinear SBE problems. However, their computational complexity is significantly increased compared to variants of the KF. In certain settings, PFs can avoid the curse of dimensionality [20]. However, they do not scale well with the dimension of the state to be estimated and are not directly amendable for distributed implementations.

Network localization for the IoT

Consider the localization of a network of IoT devices that consists of N_{a} agents (with index set $\mathcal{N}_{\mathrm{a}} = \{1,2,\ldots, N_{\mathrm{a}}\}$) and N_{b} anchors (with index set $\mathcal{N}_{\mathrm{b}} = \{N_{\mathrm{a}}+1, N_{\mathrm{a}}+2,\ldots,N_{\mathrm{a}}+N_{\mathrm{b}}\}$). Let $\mathbf{x}_i^{(n)} \in \mathbb{R}^D$ be the positional state of the node $i \in \{1,2,\ldots, N_{\mathrm{a}}+N_{\mathrm{b}}\}$. The states of all nodes are represented by the joint state vector $\mathbf{x}^{(n)} \triangleq [\mathbf{x}_1^{(n)\mathrm{T}}, \mathbf{x}_2^{(n)\mathrm{T}}, \ldots, \mathbf{x}_{N_{\mathrm{a}}+N_{\mathrm{b}}}^{(n)\mathrm{T}}]^{\mathrm{T}}$. At time n, agent $i \in \mathcal{N}_{\mathrm{a}}$ is able to communicate and perform an internode measurement $z_{ij}^{(n)}$ with nodes j in its neighbor set $\mathcal{A}_i^{(n)}$. For anchors $i \in \mathcal{N}_{\mathrm{b}}$, the neighbor set is empty, i.e., $\mathcal{A}_i^{(n)} = \emptyset$. Agent communication is symmetric, i.e., for $i,j \in \mathcal{N}_{\mathrm{a}}$, $j \in \mathcal{A}_i^{(n)}$ implies $i \in \mathcal{A}_j^{(n)}$. All measurements performed by all agents $i \in \mathcal{N}_{\mathrm{a}}$ at time n are summarized in the joint measurement vector $z^{(n)}$. Every agents aims to calculate an estimate $\hat{x}_i^{(n)}$ of $\mathbf{x}_i^{(n)}$ from all available measurements $z^{(1:n)}$ collected up to time n.

For node i at time n, the relationship of the current state vector $\mathbf{x}_i^{(n)}$ with the previous state vector $\mathbf{x}_i^{(n-1)}$ is given by the state-evolution model

$$\mathbf{x}_i^{(n)} = \boldsymbol{a}_i(\mathbf{x}_i^{(n-1)}, \mathbf{c}_i^{(n)}; \boldsymbol{u}_i^{(n)}) \tag{13}$$

where the state-evolution noise vector $\mathbf{c}_i^{(n)}$ is assumed independent across n and i. Note that the PDF $f(\boldsymbol{c}_i^{(n)})$ can be different for distinct time steps n and agent indexes i. In particular, for anchors $i \in \mathcal{N}_b$ it is assumed that $f(\boldsymbol{c}_i^{(n)}) = \delta(\boldsymbol{c}_i^{(n)})$, i.e., $\boldsymbol{c}_i^{(n)}$ is deterministic and equal to zero. From the state-evolution model (13) one can directly obtain the state-evolution function $f\left(\boldsymbol{x}_i^{(n)} \middle| \boldsymbol{x}_i^{(n-1)}; \boldsymbol{u}_i^{(n)}\right)$. At $n = 0$, the prior PDF of the joint state vector can be expressed as $f(\boldsymbol{x}^{(0)}) = \prod_{i=1}^{N_\mathrm{a}+N_\mathrm{b}} f(\boldsymbol{x}_i^{(0)})$. In particular, anchors $i \in \mathcal{N}_\mathrm{b}$ have perfect knowledge of their state, i.e., their prior PDFs are given by $f(\boldsymbol{x}_i^{(0)}) = \delta(\boldsymbol{x}_i^{(0)} - \tilde{\boldsymbol{x}}_i^{(0)})$ where $\tilde{\boldsymbol{x}}_i^{(0)}$ is the true state. Furthermore, agents have uninformative prior information $f(\boldsymbol{x}_i^{(0)})$ that is assumed known. For $n > 0$, the joint prior PDF, can be expressed as

$$\begin{aligned} f(\boldsymbol{x}^{(0:n)};\boldsymbol{u}^{(1;n)}) &= f(\boldsymbol{x}^{(0)}) \prod_{k=1}^{n} f\left(\boldsymbol{x}^{(k)} \middle| \boldsymbol{x}^{(k-1)}; \boldsymbol{u}^{(k)}\right) \\ &= \prod_{i=1}^{N_\mathrm{a}+N_\mathrm{b}} f(\boldsymbol{x}_i^{(0)}) \prod_{k=1}^{n} f\left(\boldsymbol{x}_i^{(k)} \middle| \boldsymbol{x}_i^{(k-1)}; \boldsymbol{u}_i^{(k)}\right). \end{aligned} \tag{14}$$

Agents $i \in \mathcal{N}_\mathrm{a}$ performs internode measurements $\mathbf{z}_{ij}^{(n)}$, $j \in \mathcal{A}_i^{(n)}$ that are related to the states $\mathbf{x}_i^{(n)}$ and $\mathbf{x}_j^{(n)}$ as

$$\mathbf{z}_{ij}^{(n)} = \boldsymbol{h}_{ij}(\mathbf{x}_i^{(n)}, \mathbf{x}_j^{(n)}, \mathbf{v}_{ij}^{(n)}), \tag{15}$$

where $\mathbf{v}_{ij}^{(n)}$ is the internode measurement noise. Note that the PDF $f(\boldsymbol{v}_{ij}^{(n)})$ can be different for distinct time steps n and agent indexes i, and is typically a function of the channel quality $\xi_{ij}^{(n)}$ (see the "Node Prioritization" section).

The measurement noise $\mathbf{v}_{ij}^{(n)}$ is assumed independent across all (i, j) pairs and all times n. From the measurement model (15), one can directly obtain the likelihood function $f\left(z_{ij}^{(n)} \middle| \boldsymbol{x}_i^{(n)}, \boldsymbol{x}_j^{(n)}\right)$. The joint likelihood function can be expressed as

$$f\left(z^{(1:n)} \middle| \boldsymbol{x}^{(1:n)}\right) = \prod_{k=1}^{n} \prod_{i=1}^{N_\mathrm{a}} \prod_{j \in \mathcal{A}_i^{(k)}} f\left(z_{ij}^{(k)} \middle| \boldsymbol{x}_i^{(k)}, \boldsymbol{x}_j^{(k)}\right). \tag{16}$$

Using Bayes' rules together with (14) and (16), the joint posterior PDF of $\mathbf{x}^{(0:n)}$ given $z^{(1:n)}$ for $n > 0$ is obtained as

$$\begin{aligned} & f\left(\boldsymbol{x}^{(0:n)} \middle| z^{(1:n)}; \boldsymbol{u}^{(1;n)}\right) \\ & \quad \propto f\left(z^{(1:n)} \middle| \boldsymbol{x}^{(1:n)}\right) f(\boldsymbol{x}^{(0:n)}; \boldsymbol{u}^{(1;n)}) \\ & \quad = f(\boldsymbol{x}^{(0)}) \prod_{k=1}^{n} f\left(\boldsymbol{x}^{(k)} \middle| \boldsymbol{x}^{(k-1)}; \boldsymbol{u}^{(k)}\right) f\left(z^{(k)} \middle| \boldsymbol{x}^{(k)}\right) \end{aligned} \tag{17}$$

$$\begin{aligned} & \quad = \prod_{i=1}^{N_a+N_b} f(\boldsymbol{x}_i^{(0)}) \prod_{k=1}^{n} f\left(\boldsymbol{x}_i^{(k)} \middle| \boldsymbol{x}_i^{(k-1)}; \boldsymbol{u}_i^{(k)}\right) \\ & \qquad \times \prod_{j \in \mathcal{A}_i^{(k)}} f\left(z_{ij}^{(k)} \middle| \boldsymbol{x}_i^{(k)}, \boldsymbol{x}_j^{(k)}\right). \end{aligned} \tag{18}$$

Remark 2

Note that the factorization of the marginal posterior in (17) has the same temporal structure as the marginal posterior in the single-node localization and navigation problem. The factor graph representing the factorization of the marginal posterior in (18) is shown in Figure 3. The spatiotemporal structure of the marginal posterior allows development of distributed-inference algorithms that are scalable both in time n and in the number of agents N_a as discussed in the next section.

Spatiotemporal fusion based on the SPA

In a network with multiple agents, state estimation is complicated by the fact that, since internode measurements are performed, the posterior distributions $f\left(\boldsymbol{x}_i^{(n)} \middle| z^{(1:n)}; \boldsymbol{u}^{(1:n)}\right)$ of agents are coupled and thus should be estimated jointly. A naive approach to joint sequential state estimation would be to only exploit the temporal structure of the joint posterior PDF $f\left(\boldsymbol{x}^{(0:n)} \middle| z^{(1:n)}; \boldsymbol{u}^{(1:n)}\right)$ in (17) to obtain a marginal posterior PDF $f\left(\boldsymbol{x}^{(n)} \middle| z^{(1:n)}; \boldsymbol{u}^{(1:n)}\right)$ by means of an algorithm presented in the "Node Localization and Navigation Algorithms" section and then calculating an estimate for the joint agent state $\boldsymbol{x}^{(n)}$. However, this approach is not scalable, as the dimension of $\boldsymbol{x}^{(n)}$ increases with the number of agents N_a. In addition, it is not amenable for a distributed implementation because it necessitates the existence of a fusion center that collects all pairwise measurements performed in the network.

Alternatively, distributed and scalable estimation can be performed by running SPA on the factor graph shown in Figure 3. In the case of a factor graph with loops, the beliefs produced by the SPA are generally only approximations of the marginal posterior PDFs and they typically suffer from overconfidence (in the sense that the uncertainty of the estimates is underestimated by their spread). Furthermore, there is no fixed order for message calculation in loopy SPA, and different orders may lead to different beliefs. This means that there is a certain freedom to design the order of messages in the development of SPA algorithms.

The message-passing rules presented next are obtained by 1) applying SPA [18] to the factor graph in Figure 3, 2) performing temporal fusion by sending messages only forward in time, and 3) performing only a single message-passing iteration in the spatial fusion step. In the temporal fusion step at agent i and time n, since messages are sent only forward in time, the messages $\alpha_{f_i}(\boldsymbol{x}_i^{(n-1)})$ are equal to the beliefs computed at $n - 1$, i.e., [18]

$$\alpha_{f_i}(\boldsymbol{x}_i^{(n-1)}) = b(\boldsymbol{x}_i^{(n-1)}). \tag{19}$$

Therefore, the messages $\beta_{f_i}(\boldsymbol{x}_i^{(n)})$ can be obtained as

$$\begin{aligned} \beta_{f_i}(\boldsymbol{x}_i^{(n)}) &= \int f\left(\boldsymbol{x}_i^{(n)} \middle| \boldsymbol{x}_i^{(n-1)}; \boldsymbol{u}_i^{(n)}\right) \alpha_{f_i}(\boldsymbol{x}_i^{(n-1)}) \mathrm{d}\boldsymbol{x}_i^{(n-1)} \\ &= \int f\left(\boldsymbol{x}_i^{(n)} \middle| \boldsymbol{x}_i^{(n-1)}; \boldsymbol{u}_i^{(n)}\right) b(\boldsymbol{x}_i^{(n-1)}) \mathrm{d}\boldsymbol{x}_i^{(n-1)}. \end{aligned} \tag{20}$$

Note that the calculation of the message $\beta_{f_i}(\boldsymbol{x}_i^{(n)})$ in the temporal fusion step is equivalent to the prediction step of SBE in (7) and its SPA interpretation in (9).

In the spatial fusion step, since only a single message-passing iteration is performed, outgoing messages $\alpha_{f_{ij}}(\boldsymbol{x}_i^{(n)}), i \in \mathcal{A}_j$ passed from variable node $\boldsymbol{x}_i$ to factor nodes f_{ij} are directly given by $\alpha_{f_{ij}}(\boldsymbol{x}_i^{(n)}) = \beta_{f_i}(\boldsymbol{x}_i^{(n)})$. Furthermore, incoming messages $\beta_{f_{ij}}(\boldsymbol{x}_i^{(n)}), j \in \mathcal{A}_i$ can be obtained as

$$\begin{aligned}\beta_{f_{ij}}(\boldsymbol{x}_i^{(n)}) &= \int f\left(z_{ij}^{(n)} \middle| \boldsymbol{x}_i^{(n)}, \boldsymbol{x}_j^{(n)}\right)\alpha_{f_{ji}}(\boldsymbol{x}_j^{(n)})\mathrm{d}\boldsymbol{x}_j^{(n)}\\ &= \int f\left(z_{ij}^{(n)} \middle| \boldsymbol{x}_i^{(n)}, \boldsymbol{x}_j^{(n)}\right)\beta_{f_i}(\boldsymbol{x}_j^{(n)})\mathrm{d}\boldsymbol{x}_j^{(n)}. \end{aligned} \tag{21}$$

Finally, the belief of an agent i at time n is calculated as

$$b(\boldsymbol{x}_i^{(n)}) \propto \beta_{f_i}(\boldsymbol{x}_i^{(n)}) \prod_{j \in \mathcal{A}_i^{(n)}} \beta_{f_{ij}}(\boldsymbol{x}_i^{(n)}). \tag{22}$$

The messages $\beta_{f_i}(\boldsymbol{x}_i^{(n)})$ in (20) and the belief $b(\boldsymbol{x}_i^{(n)})$ in (22) are PDFs, i.e., they integrate to one. The belief $b(\boldsymbol{x}_i^{(n)}) \approx f\left(\boldsymbol{x}_i^{(n)} \middle| \boldsymbol{z}^{(1:n)}; \boldsymbol{u}^{(1:n)}\right)$ can now be used to calculate an estimate $\hat{\boldsymbol{x}}_i^{(n)}$ of the positional state of agent i at time n. Note that for anchors $i \in \mathcal{N}_{\mathrm{b}}$, the belief and the messages are given by $b(\boldsymbol{x}_i^{(n)}) = \alpha_{f_i}(\boldsymbol{x}_i^{(n)}) = \beta_{f_i}(\boldsymbol{x}_i^{(n)}) = \delta(\boldsymbol{x}_i^{(n)} - \tilde{\boldsymbol{x}}_i^{(n)})$ and $\mathcal{A}_i^{(n)} = \emptyset$ for all n.

Contrary to SBE, which only exploits the temporal structure of the estimation problem, loopy SPA performed on the factor graph in Figure 3 also exploits spatial structure. Increasing the number of agents leads to additional variable nodes in the factor graph but not to a higher dimension of the exchanged SPA messages. Therefore, the curse of dimensionality in time n and in network size $N_{\mathrm{a}} + N_{\mathrm{b}}$ is avoided. As will be discussed next, message passing according to (19)–(22) nearly automatically yields to a distributed implementation.

Distributed network-localization algorithms

We now present a framework for designing network-localization algorithms that is based on a reformulation of SPA for spatiotemporal fusion (19)–(22) as local instances of SBE performed on each agent [5], [6]. Within this framework, spatiotemporal fusion is possible in a scalable and distributed way by directly applying arbitrary existing algorithms based on SBE,

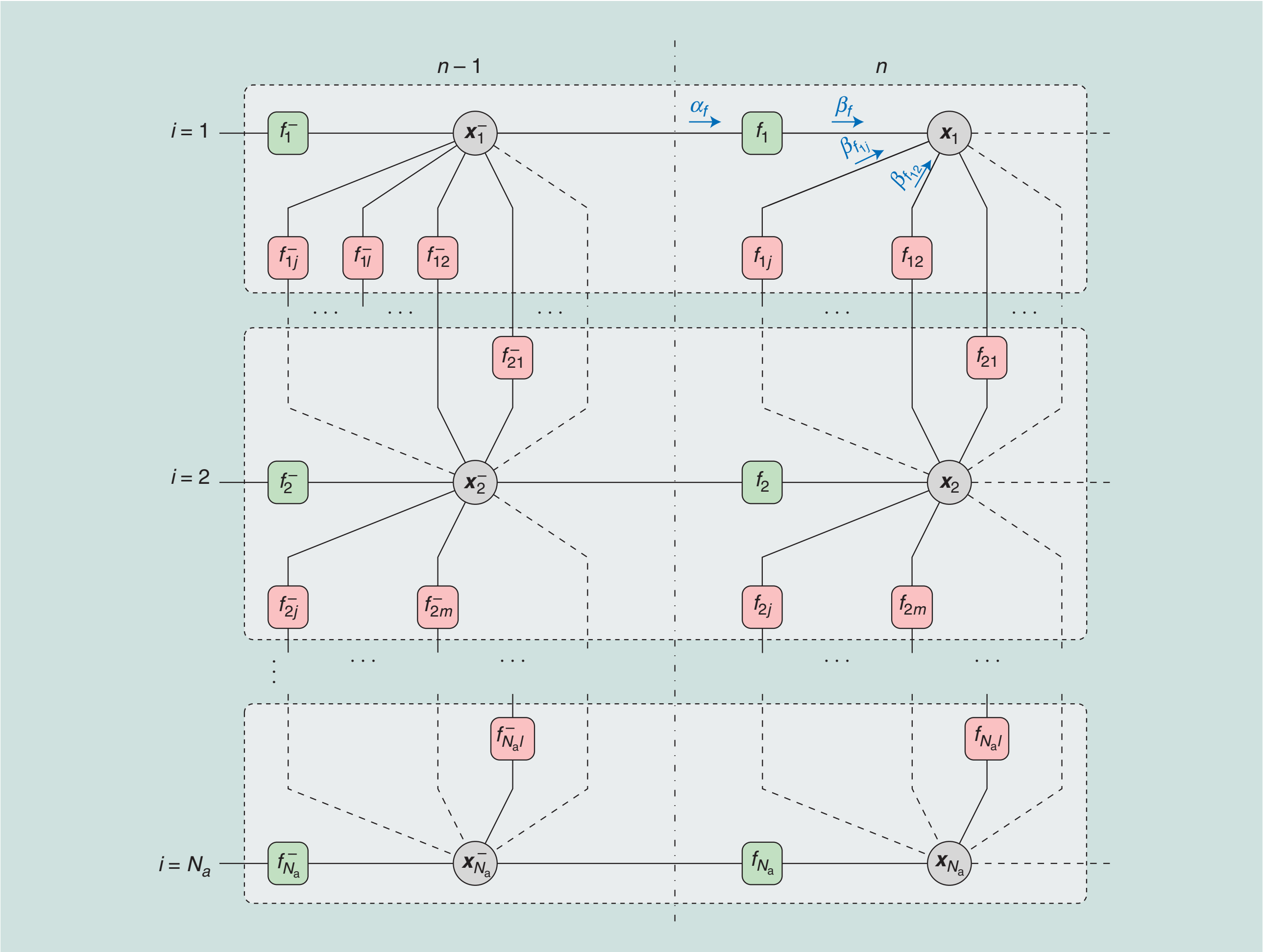


FIGURE 3. Two time steps of the factor graph for network localization corresponding to the factorizes (18). Nodes in green represent factors related to the state-evolution function, nodes in red represent factors related to the likelihood function, while SPA messages are in blue. The following short notations are used: $\boldsymbol{x}_i^- \triangleq \boldsymbol{x}_i^{(n-1)}$, $\boldsymbol{x}_i \triangleq \boldsymbol{x}_i^{(n)}$, $f_i \triangleq f\left(\boldsymbol{x}_i^{(n)} \,|\, \boldsymbol{x}_i^{(n-1)}; \boldsymbol{u}_i^{(n)}\right)$, $f_i^- \triangleq f\left(\boldsymbol{x}_i^{(n-1)} \,|\, \boldsymbol{x}_i^{(n-2)}; \boldsymbol{u}_i^{(n-1)}\right)$, $f_{ij} \triangleq f\left(\boldsymbol{z}_{ij}^{(n)} \,|\, \boldsymbol{x}_i^{(n)}, \boldsymbol{x}_j^{(n)}\right)$, $f_{ij}^- \triangleq f\left(\boldsymbol{z}_{ij}^{(n-1)} \,|\, \boldsymbol{x}_i^{(n-1)}, \boldsymbol{x}_j^{(n-1)}\right)$, $\alpha_{f_i} \triangleq \alpha_{f_i}(\boldsymbol{x}_i^{(n-1)})$, $\beta_{f_i} \triangleq \beta_{f_i}(\boldsymbol{x}_i^{(n)})$, and $\beta_{f_{ij}} \triangleq \beta_{f_{ij}}(\boldsymbol{x}_i^{(n)})$.

such as those reviewed in the “Node Localization and Navigation Algorithms” section.

Consider the spatiotemporal fusion at agent i, and introduce the augmented state vector $\bar{\mathbf{x}}_i^{(n)}$ and the augmented measurement $\bar{\mathbf{z}}_i^{(n)}$ as

$$\bar{\mathbf{x}}_i^{(n)} = [\mathbf{x}_j^{(n)}]_{j\in\{i\}\cup\mathcal{A}_i^{(n)}} \quad \bar{\mathbf{z}}_i^{(n)} = [\mathbf{z}_{ij}^{(n)}]_{j\in\mathcal{A}_i^{(n)}}.$$

Moreover, the belief $b(\bar{\boldsymbol{x}}_i^{(n)})$ of $\bar{\mathbf{x}}_i^{(n)}$ is introduced as

$$b(\bar{\boldsymbol{x}}_i^{(n)}) \propto f\left(\bar{z}_i^{(n)}\middle|\bar{\boldsymbol{x}}_i^{(n)}\right) f(\bar{\boldsymbol{x}}_i^{(n)}), \tag{23}$$

where the “prior” $f(\bar{\boldsymbol{x}}_i^{(n)})$ and the “likelihood” function $f\left(\bar{z}_i^{(n)}\middle|\bar{\boldsymbol{x}}_i^{(n)}\right)$ are given by

$$f(\bar{\boldsymbol{x}}_i^{(n)}) = \prod_{j\in\{i\}\cup\mathcal{A}_i^{(n)}} \beta_{f_j}(\boldsymbol{x}_j^{(n)}) \tag{24}$$

$$f\left(\bar{z}_i^{(n)}\middle|\bar{\boldsymbol{x}}_i^{(n)}\right) = \prod_{j\in\mathcal{A}_i^{(n)}} f\left(z_{ij}^{(n)}\middle|\boldsymbol{x}_i^{(n)}, \boldsymbol{x}_j^{(n)}\right). \tag{25}$$

Note that here, with an abuse of notation, control inputs $\boldsymbol{u}_i^{(k)}$ and measurements $z_{ij}^{(k)}$ from previous time steps $k \in \{1, 2, \ldots, n-1\}$ are avoided. The expression (23) has the same form as the update step of SBE in (8).

By plugging (21) into (22) and subsequently swapping the order of multiplication and integration, (22) becomes

$$b(\boldsymbol{x}_i^{(n)}) = \int b(\bar{\boldsymbol{x}}_i^{(n)})\mathrm{d}\bar{\boldsymbol{x}}_{\sim i}^{(n)}, \tag{26}$$

where $\bar{\boldsymbol{x}}_{\sim i}^{(n)}$ is the vector obtained by removing $\boldsymbol{x}_i^{(n)}$. from $\bar{\boldsymbol{x}}_i^{(n)}$.

Equations (23) and (26) indicate that $b(\boldsymbol{x}_i^{(n)})$ can be obtained via an update step (8) followed by marginalization. This observation motivates the following three steps at each agent $i \in \mathcal{N}_\mathrm{a}$ to perform spatiotemporal fusion by means of SPA.

- *Step 1*: *Local Prediction and Information Exchange*. Agent i calculates $\beta_{f_i}(\boldsymbol{x}_i^{(n)})$ locally according to (20) which is equivalent to the prediction step in (7). [The prediction step of any algorithm based on SBE, such as those presented in the “Message-Passing Interpretation of SBE” section, can be used to calculate $\beta_{f_i}(\boldsymbol{x}_i^{(n)})$]. Then each agent broadcasts $\beta_{f_i}(\boldsymbol{x}_i^{(n)})$ and receives $\beta_{f_j}(\boldsymbol{x}_j^{(n)})$ from its neighbors $j \in \mathcal{A}_i^{(n)}$ so that $f(\bar{\boldsymbol{x}}_i^{(n)})$ in (24) becomes available at agent i.
- *Step 2*: *Measurement Phase and State Update*. Agent i cooperates with its neighbors $j \in \mathcal{A}_i^{(n)}$ to acquire internode measurements $z_{ij}^{(n)}$. Now the likelihood function $f\left(\bar{z}_i^{(n)}\middle|\bar{\boldsymbol{x}}_i^{(n)}\right)$ in (25) is available at agent i and the belief $b(\bar{\boldsymbol{x}}_i^{(n)})$ of $\bar{\mathbf{x}}_i^{(n)}$ can be calculated locally by performing the update step in (23). Note that the update step of any algorithm based on SBE such as those presented in the “Message-Passing Interpretation of SBE” section can be used to calculate $b(\bar{\boldsymbol{x}}_i^{(n)})$.
- *Step 3*: *Marginalization*. In this step, agent i computes the belief $b(\boldsymbol{x}_i^{(n)})$ from $b(\bar{\boldsymbol{x}}_i^{(n)})$. This typically incurs no computational overhead. For example, if $b(\bar{\boldsymbol{x}}_i^{(n)})$ is represented by the mean vector $\bar{\boldsymbol{\mu}}_i^{(n)}$ and the covariance matrix $\bar{\boldsymbol{\Sigma}}_i^{(n)}$, then the mean vector $\boldsymbol{\mu}_i^{(n)}$ and the covariance matrix $\boldsymbol{\Sigma}_i^{(n)}$ related to $b(\boldsymbol{x}_i^{(n)})$ can be directly extracted from $\bar{\boldsymbol{\mu}}_i^{(n)}$ and $\bar{\boldsymbol{\Sigma}}_i^{(n)}$, respectively. In case a particle representation $\{(\bar{\boldsymbol{x}}_{i,l}^{(n)}, w_{i,l}^{(n)})\}_{l=1}^{L}$ of the belief $b(\bar{\boldsymbol{x}}_i^{(n)})$ is available, a particle representation $\{(\boldsymbol{x}_{i,l}^{(n)}, w_{i,l}^{(n)})\}_{l=1}^{L}$ of the belief $b(\boldsymbol{x}_i^{(n)})$ can be obtained by discarding from the particles $\bar{\boldsymbol{x}}_{i,l}^{(n)}$ all subvectors $\boldsymbol{x}_{j,l}^{(n)}$ with $j \neq i$.

Note that the belief $b(\boldsymbol{x}_i^{(n)})$ can be calculated by only communicating with neighboring agents in the network. For accurate localization and navigation of an agent $i \in \mathcal{N}_\mathrm{a}$, typically only a small number of neighbors $\left|\mathcal{A}_i^{(n)}\right|$ are necessary. Therefore, the communication cost related to the information exchange in Step 1 as well as the computation cost related to calculating the beliefs $b(\bar{\boldsymbol{x}}_i^{(n)})$ remain feasible. More importantly, for a single agent $i \in \mathcal{N}_\mathrm{a}$, these costs only depend on the number of neighbors $\left|\mathcal{A}_j^{(n)}\right|$ but not on the network size $N_\mathrm{a} + N_\mathrm{b}$. An attractive property of calculating $b(\boldsymbol{x}_i^{(n)})$ by means of Steps 1–3 is that existing techniques for single-node localization and navigation can be directly leveraged for scalable and distributed network localization. Note that SP belief propagation (SPBP) [5] and the network-localization algorithm in [6] have been developed according to Steps 1–3.

Efficient network operation

Network-operation strategies [21], [26] are indispensable for efficient localization and navigation in IoT scenarios. The network-operation strategies presented in this article focus on the coordination of measurements provided by range measurement units (RMUs), i.e., the measurement model in (15) is

$$\mathsf{z}_{ij}^{(n)} = \left\|\mathbf{x}_i^{(n)} - \mathbf{x}_j^{(n)}\right\| + \mathsf{v}_{ij}^{(n)}.$$

The performance of RMUs such as ultrawideband (UWB) radios is often limited by the fact that [16], [27], [28]:

1) Agents often make measurements with nodes with low link quality or poor geometry.
2) Different agents, which simultaneously transmit ranging signals, interfere with each other.

To address these issues, node-activation strategies to reduce interference and node prioritization strategies to allocate resources to measurements with neighbor nodes can be employed. A flowchart that visualizes the interaction of node activation, node prioritization, network localization, and the RMU is shown in Figure 4.

Note that, in what follows, the inverse Fisher information matrix [3] is referred to as an *error matrix*. In particular, all strategies developed in this article rely either on the individual error matrices $\boldsymbol{Q}_i^{(n)}$ related to the positions $\mathbf{p}_i^{(n)}$ of the agents $i \in \mathcal{N}_\mathrm{a}$ or on the joint error matrix $\boldsymbol{Q}^{(n)}$ related to the individual positions of all agents, as defined in [24]. These error matrices are not accessible in real-world localization systems as they rely on the knowledge of true positions. For this reason, in an implementation of the presented node-operation strategies [16], these error matrices are approximated by the corresponding covariance matrices, which can be provided by network-localization algorithms.

Node activation

Node-activation strategies enable a significant reduction of packet collisions and localization errors in the network. The goal of node-activation strategies is to determine a set of nodes that are permitted to make range measurements so that packet collisions are avoided and the localization error reduction of the network is maximized. In what follows, we discuss centralized and distributed strategies for node activation.

Centralized node activation

If agent i is selected to make internode measurements with its neighbors at time n, the error evolution relationship is given by [24]

$$\boldsymbol{Q}^{(n+1)} = \left((\boldsymbol{Q}^{(n)})^{-1} + \sum_{j \in \mathcal{A}_i^{(n)}} \boldsymbol{S}_{ij}^{(n)}\right)^{-1} + \Delta^{(n+1)},$$

where $\boldsymbol{S}_{ij}^{(n)}$ denotes the information matrix corresponding to the measurement $z_{ij}^{(n+1)}$, and $\Delta^{(n+1)}$ denotes the matrix corresponding to the error introduced in the temporal cooperation step. Note that $\boldsymbol{S}_{ij}^{(n)}$ also depends on the amount of resources y_{ij} allocated to the measurements link (i, j) that can be determined by node prioritization discussed in the "Distributed node Activation" section [24].

Centralized node activation can be performed by calculating the agent index i_n that is optimum, in the sense that the localization error reduction of the network is maximized. The optimum index can be obtained as follows:

$$i_n = \max_{i \in \mathcal{N}_{\mathrm{a}}} \mathrm{tr}\left((\boldsymbol{Q}^{(n)})^{-1} + \sum_{j \in \mathcal{A}_i^{(n)}} \boldsymbol{S}_{ij}^{(n)}\right)^{-1}. \tag{27}$$

This node-activation strategy is one-step optimal because the active node is selected such that the localization error at time $n + 1$ is minimized. Alternatively, one can also try to activate nodes so that the average error over multiple time instants is minimized. Such a problem can be solved through dynamic programming, but the computational complexity increases rapidly with the number of time steps. Note that the evaluation of (27) relies on the joint error matrix $\boldsymbol{Q}^{(n)}$. The centralized node-activation strategy is thus not scalable with the network size since it necessitates a central controller that collects the information of all the agents in the network. For this reason, for large-scale NLN, distributed node-activation strategies are needed.

Distributed node activation

Consider the case in which the activation set may consist of multiple agents. In particular, at time n every agent i tries to make distance measurements with its neighbors $j \in \mathcal{A}_i^{(n)}$ with a certain channel access probability $\zeta_i^{(n)}$. The one-step optimization problem that minimizes the localization error over the channel access probabilities $\zeta_i^{(n)}$ is given by

$$\begin{aligned} \mathcal{P}_{\mathrm{NA}}^{(n)}:\quad & \underset{\{\zeta_i^{(n)}\}_{i \in \mathcal{N}_{\mathrm{a}}}}{\text{minimize}} \quad \mathbb{E}\left\{\mathrm{tr}\{\mathbf{Q}^{(n+1)}\}\middle|\,\mathbf{Q}^{(n)}\right\} \\ & \text{subject to} \quad 0 \le \zeta_i^{(n)} \le 1, \quad i \in \mathcal{N}_{\mathrm{a}}, \end{aligned}$$

where the expectation in the objective function is over the randomness in the channel access event for all the agents. In [26], the optimal channel access probabilities $\zeta_i^{(n)}, i \in \mathcal{N}_{\mathrm{a}}$ resulting from $\mathcal{P}_{\mathrm{NA}}^{(n)}$ can be obtained as

$$\zeta_i^{(n)} = \begin{cases} 1, & \text{if } \mathcal{X}_i^{(n)} > \sum\limits_{j \in \mathcal{A}_i^{(n)} \cup \{i\}} \Delta_j^{(n)} \\ 0, & \text{otherwise} \end{cases}, \tag{28}$$

where $\mathcal{X}_i^{(n)}$ denotes the expected error reduction of agent i, if it is activated and successful, makes range measurements with its neighbors $\mathcal{A}_i^{(n)}$ and $\Delta_j^{(n)}$, and denotes the error increase of the agents in the subnetwork $\mathcal{A}_i^{(n)} \cup \{i\}$ during the time range measurements are performed. Note that $\mathcal{X}_i^{(n)}$ and $\Delta_j^{(n)}$ are functions of $\boldsymbol{Q}_i^{(n)}$ and $\boldsymbol{Q}_j^{(n)}$, $j \in \mathcal{A}_i \cup \{i\}$, respectively.

Remark 3

This optimal strategy $\mathcal{P}_{\mathrm{NA}}^{(n)}$ leads to a nonrandom node activation in the sense that an agent accesses the channel either with probability one or with probability zero. Moreover, the optimal strategy is distributed because for agent i, $\chi_i^{(n)}$ and $\Delta_j^{(n)}$ can be determined or accurately approximated using information that is either locally available or has been received from neighboring nodes $j \in \mathcal{A}_i^{(n)}$. Unlike the setting in the centralized node activation, the distributed strategy may activate multiple nodes at the same time and cause packet collisions. The possibility of these collision events can be reduced by incorporating channel sensing in the presented activation strategy. This results in the distributed node-activation strategy presented in Algorithm 1 that has been successfully verified on-the-field with the Peregrine system for 3-D NLN.

Node prioritization

Node-prioritization strategies provide a desirable tradeoff between resource consumption and localization accuracy. In

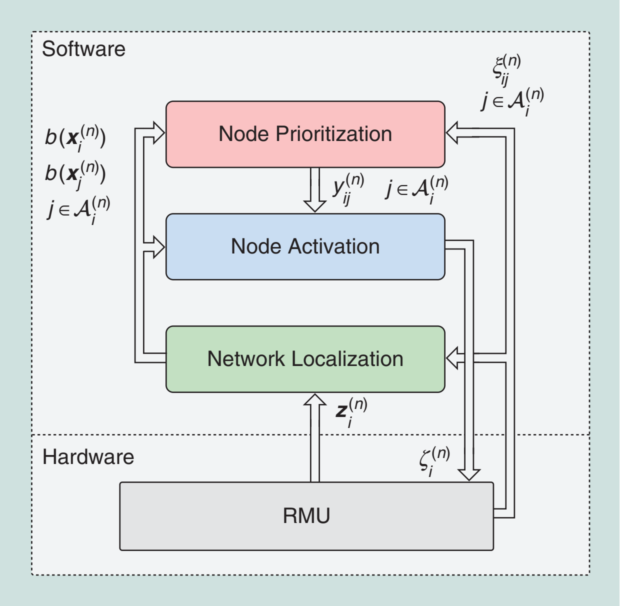


FIGURE 4. A flowchart showing the interaction of node activation, node prioritization, network localization, and the RMU.

what follows, we again discuss centralized and distributed strategies for node prioritization.

Centralized node prioritization

For time $n+1$, the error matrix $\boldsymbol{Q}^{(n+1)}$ can be obtained [24] as

$$\boldsymbol{Q}^{(n+1)} = \Big((\boldsymbol{Q}^{(n)})^{-1} + \sum_{(i,j)\in\mathcal{E}^{(n)}} y_{ij}^{(n)} \xi_{ij}^{(n)} \boldsymbol{u}_{ij}^{(n)} \boldsymbol{u}_{ij}^{(n)\mathrm{T}} \Big)^{-1} + \Delta^{(n+1)},$$

where $\mathcal{E}^{(n)} = \{(i,j) \;:\; i \in \mathcal{N}_{\mathrm{a}}, j \in \mathcal{A}_i^{(n)}, i > j\}$ is the set of candidate measurement link pairs, $y_{ij}^{(n)}$ is the amount of resources allocated to the measurement link pair (i,j), $\xi_{ij}^{(n)}$ represents the channel quality between nodes i and j, and $\boldsymbol{u}_{ij}^{(n)}$ is given in [21] and the "Spatiotemporal Fusion Based on the SPA" section and depends on the relative positions of nodes i and j. Furthermore, $y_{ij}^{(n)}$ are the variables to be optimized. As a special case, if only node i is activated, $\mathcal{E}^{(n)} = \{(i,j) : j \in \mathcal{A}_i^{(n)}\}$.

Now the following optimization problem for centralized node prioritization can be introduced

$$\begin{aligned} \mathcal{P}_{\mathrm{NP-C}}^{(n)} \;:\; & \underset{\{y_{ij}^{(n)}\}_{(i,j)\in\mathcal{E}^{(n)}}}{\text{minimize}} \quad \mathrm{tr}(\boldsymbol{Q}^{(n+1)}) \\ & \text{subject to} \quad l_k\big(\{y_{ij}^{(n)}\}_{(i,j)\in\mathcal{E}^{(n)}}\big) \leqslant 0,\; k \in \mathcal{L}, \end{aligned}$$

where $\mathcal{L}$ is the set of linear constraints $l_k(\cdot)$. Due to the special structure of $\boldsymbol{Q}^{(n+1)}$, $\mathcal{P}_{\mathrm{NP-C}}^{(n)}$ can be transformed to the following semidefinite program (SDP):

$$\begin{aligned} & \underset{\boldsymbol{M},\{y_{ij}^{(n)}\}_{(i,j)\in\mathcal{E}^{(n)}}}{\text{minimize}} \quad \mathrm{tr}(\boldsymbol{M}) \\ & \text{subject to} \quad \begin{bmatrix} \boldsymbol{M} & \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{J}^{(n)} \end{bmatrix} \succcurlyeq 0 \\ & \qquad \boldsymbol{J}^{(n)} = (\boldsymbol{Q}^{(n)})^{-1} \\ & \qquad\qquad + \sum_{(i,j)\in\mathcal{E}^{(n)}} y_{ij}^{(n)} \xi_{i,j}^{(n)} \boldsymbol{u}_{ij}^{(n)} \boldsymbol{u}_{ij}^{(n)\mathrm{T}} \\ & \qquad l_k\big(\{y_{ij}^{(n)}\}_{(i,j)\in\mathcal{E}^{(n)}}\big) \leqslant 0, \quad k \in \mathcal{L}, \end{aligned}$$

where $\boldsymbol{M}$ is an auxiliary matrix for the SDP formulation [21]. Convex optimization engines [29] can be used to solve the SDP in the above equation. Note that similarly to the node-activation problem, solving the node prioritization problem $\mathcal{P}_{\mathrm{NP-C}}^{(n)}$ requires obtaining the estimates of $\boldsymbol{Q}^{(n)}$, $y_{ij}^{(n)}$, $\xi_{i,j}^{(n)}$, and $\boldsymbol{U}_{ij}^{(n)}$ for the solution of this SDP. A central controller is needed to collect such information. Moreover, the computational complexity of this SDP largely depends on the dimension of $\boldsymbol{Q}^{(n)}$, which is a $DN_{\mathrm{a}} \times DN_{\mathrm{a}}$ matrix. For these reasons, centralized node prioritization does not scale with the size of the network.

Algorithm 1. Distributed node-activation strategy.

```
for all $i \in \mathcal{N}_a$ do
    Agent $i$ listens to the channel;
    if the channel is busy then
        Wait for a certain amount of time;
    else
        Determine the access probability $\zeta_i^{(n)}$ from (28);
        if $\zeta_i^{(n)} = 1$ then
            Access the channel and perform internode measurements;
        end if
    end if
    Broadcast $\Delta_i^{(n)}$;
end for
```

Distributed node prioritization

Though the centralized formulation can provide better localization performance, in large networks it incurs in extensive communication overhead and computational complexity. For this reason, fully distributed and thus scalable variants are more amenable in practice.

The error matrix for the position of agent i is the ith diagonal $D \times D$ block of $\boldsymbol{Q}^{(n+1)}$, denoted by $[\boldsymbol{Q}^{(n+1)}]_i$. This error matrix depends on the geometry of the network and the accuracies of all internode measurements. Therefore, directly optimizing this error matrix does not lead to distributed implementation. An approximation of $[\boldsymbol{Q}^{(n+1)}]_i$ that involves only local parameters can be introduced as follows:

$$[\boldsymbol{Q}^{(n+1)}]_i \approx \tilde{\boldsymbol{Q}}_i^{(n+1)}, \tag{29}$$

where

$$\tilde{\boldsymbol{Q}}_i^{(n+1)} = \Big(([\boldsymbol{Q}^{(n)}]_i)^{-1} + \sum_{j\in\mathcal{A}_i} y_{ij}^{(n)} \varrho_{ij}^{(n)} \boldsymbol{v}_{ij}^{(n)} [\boldsymbol{v}_{ij}^{(n)}]^{\mathrm{T}} \Big)^{-1} + [\Delta^{(n+1)}]_i.$$

In this expression, $\varrho_{ij}^{(n)}$ is given by

$$\varrho_{ij}^{(n)} = \begin{cases} \xi_{ij}^{(n)}, & \text{if } j \in \mathcal{A}_i^{(n)} \cap \mathcal{N}_{\mathrm{b}} \\ \dfrac{\xi_{ij}^{(n)}}{1 + y_{ij}^{(n)} \mathrm{tr}\big(\boldsymbol{u}_{ij}^{(n)} [\boldsymbol{u}_{ij}^{(n)}]^{\mathrm{T}} [\boldsymbol{Q}^{(n)}]_j\big)} & \text{if } j \in \mathcal{A}_i^{(n)} \cap \mathcal{N}_{\mathrm{a}}\text{'} \end{cases}$$

and $\boldsymbol{v}_{ij}^{(n)} \in \mathbb{R}^D$ is a unit vector representing the direction between node i and j. Note that $\tilde{\boldsymbol{Q}}_i^{(n+1)}$ involves $\varrho_{ij}^{(n)}$, $\boldsymbol{v}_{ij}^{(n)}$, and $y_{ij}^{(n)}$ for $j \in \mathcal{A}_i^{(n)}$, which are either locally available at agent i or can be received by communicating with neighboring nodes $j \in \mathcal{A}_i^{(n)}$.

Using $\mathrm{tr}(\tilde{\boldsymbol{Q}}_i^{(n+1)})$ as the objective function, a distributed node-prioritization problem is formulated as

$$\begin{aligned} \mathcal{P}_{i,\mathrm{NP-D}}^{(n)}: \quad & \underset{\{y_{ij}^{(n)}\}_{j\in\mathcal{N}_{\mathrm{a}}\cup\mathcal{N}_{\mathrm{b}}\setminus\{i\}}}{\text{minimize}} \quad \mathrm{tr}(\tilde{\boldsymbol{Q}}_i^{(n+1)}) \\ & \text{subject to} \quad l_{ik}\Big(\{y_{ij}^{(n)}\}_{j\in\mathcal{A}_i}\Big) \leq 0, k \in \mathcal{L}_i, \end{aligned} \tag{30}$$

where $l_{ik}(\cdot)$ are linear constraints [21]. It can be shown that $\mathcal{P}_{i,\mathrm{NP-D}}^{(n)}$ is a convex problem by performing the same steps as in [21]. Moreover, for a general D, one can show that $\mathcal{P}_{i,\mathrm{NP-D}}^{(n)}$ is an SDP. For $D = 2$, $\tilde{\boldsymbol{Q}}_i^{(n+1)}$ is a 2×2 matrix and $\mathrm{tr}(\tilde{\boldsymbol{Q}}_i^{(n+1)})$ has a simpler explicit expression as a function of $y_{ij}^{(n)}$. As a consequence, $\mathcal{P}_{i,\mathrm{NP-D}}^{(n)}$ can be further transformed into a second-order cone program [22], [29].

So far, we have discussed node prioritization for cooperative IoT networks. In noncooperative scenarios where agents only perform agent-anchor range measurements, the approximation (29) becomes an equality and the error matrix for agent i is

$$[\boldsymbol{Q}^{(n+1)}]_i = \Big(([\boldsymbol{Q}^{(n)}]_i)^{-1} + \sum_{j \in \mathcal{N}_\mathrm{b}} y_{ij}^{(n)} \xi_{ij}^{(n)} \boldsymbol{v}_{ij}^{(n)} [\boldsymbol{v}_{ij}^{(n)}]^\mathrm{T}\Big)^{-1} + [\Delta^{(n+1)}]_i \,.$$

Remark 4

The node-prioritization problem in noncooperative scenarios is a special case of $\mathcal{P}_{i,\mathrm{NP-D}}^{(n)}$ that can be solved even more efficiently by using geometric optimization methods [15]. Furthermore, if the constraint (30) can be expressed as follows

$$\sum_{j \in \mathcal{N}_\mathrm{b}} y_{ij}^{(n)} \leqslant R_\mathrm{tot} \qquad \text{with } y_{ij}^{(n)} \geqslant 0,\ j \in \mathcal{N}_\mathrm{b}\,,$$

the optimal solution is demonstrated to have a sparsity property. Note that here R_tot is the total amount of available resources. In particular, the optimal set of measurements can be performed with at most $D(D+1)/2$ anchors. This sparsity property provides a theoretical basis for reducing measurement links in localization networks.

Case study

In this section, we demonstrate the performance benefits of cooperation among devices and multisensor fusion in a large-scale IoT network using simulated measurements. Some of the presented algorithms have also been evaluated in the real-time localization system called *Peregrine* [16]. (A video that demonstrates how this system operates and the performance advantages related to the proposed algorithms is available online at http://winslab.lids.mit.edu/nln-technology-readiness.mp4.)

Scenario

An IoT network that consists of 512 mobile agents and 27 anchors is considered. The anchors form an equally spaced 3-D grid, where possible coordinate values on each axis in 3-D space are $\{-60, 0, 60\}$ m. Mobile agents are equipped with an inertial measurement unit (IMU) and an RMU, and they infer navigation information every $\Delta T = 0.05$ s. This scenario is inspired by a swarm of micro unmanned aerial vehicles (UAVs) that operate in a large building such as a stadium or warehouse.

The state $\mathbf{x}_i^{(n)}$, of agent $i \in \mathcal{N}_\mathrm{a}$ consists of its position $\mathbf{p}_i^{(n)} = [p_{i,1}^{(n)}\ p_{i,2}^{(n)}\ p_{i,3}^{(n)}]^\mathrm{T} \in \mathbb{R}^3$, velocity $\dot{\mathbf{p}}_i^{(n)} \in \mathbb{R}^3$, and its orientation represented by an unit quaternion $\mathbf{q}_i^{(n)} \in \mathbb{R}^4$. The initial states $\mathbf{x}_i^{(1)}, i \in \mathcal{N}_\mathrm{a}$ are chosen as follows. The initial positions $\mathbf{p}_i^{(1)}$ are sampled from the PDF that is uniform on the 3-D cube $\mathcal{R} = [-60, 60]\,\mathrm{m} \times [-60, 60]\,\mathrm{m} \times [-60, 60]\,\mathrm{m}$; the initial velocity is set to $\dot{\mathbf{p}}_i^{(1)} = \mathbf{0}$ m/s, and the initial quaternion is set to $\mathbf{q}_i^{(1)} = [1\ 0\ 0\ 0]^\mathrm{T}$. The trajectories of the agents are generated randomly. The parts of the trajectories that are related to the substates $\mathbf{s}_i^{(n)} := [[\mathbf{p}_i^{(n)}]^\mathrm{T}\ [\dot{\mathbf{p}}_i^{(n)}]^\mathrm{T}]^\mathrm{T}$ are generated by means of a constant velocity motion model [17]. More specifically, at time n the new substate $\mathbf{s}_i^{(n)}$ of agent $i \in \mathcal{N}_\mathrm{a}$ is obtained from $\mathbf{s}_i^{(n-1)}$ as

$$\mathbf{s}_i^{(n)} = \boldsymbol{A}\mathbf{s}_i^{(n-1)} + \boldsymbol{C}\mathbf{g}_i^{(n)},$$

where matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ are given as in [17] and $\mathbf{g}_i^{(n)} \in \mathbb{R}^3$ is the acceleration vector in the global reference frame.

Vector $\mathbf{g}_i^{(n)}$ consists of the random driving noise $\mathbf{r}_i^{(n)}$ and the drag force $\mathbf{f}_i^{(n)}$, i.e., $\mathbf{g}_i^{(n)} = \mathbf{r}_i^{(n)} + \mathbf{f}_i^{(n)}$. In particular, $\mathbf{r}_i^{(n)}$ is a zero-mean Gaussian random vector, i.e., $\mathbf{r}_i^{(n)} \sim \mathcal{N}(\mathbf{0}, \sigma_\mathbf{r}^2 \boldsymbol{I}_3)$ and the drag force is given by $\mathbf{f}_i^{(n)} = [f_{1,i}^{(n)}\ f_{2,i}^{(n)}\ f_{3,i}^{(n)}]^\mathrm{T}$ with elements $f_{k,i}^{(n)\mathrm{T}} = -\gamma_f \dot{\mathrm{p}}_{k,i}^{(n-1)} \big|\dot{\mathrm{p}}_{k,i}^{(n-1)}\big|$, $k \in \{1, 2, 3\}$. The drag force is introduced to limit the speed of the agents. The following parameters are used: $\sigma_\mathbf{r} = 4.0\ \mathrm{m/s}^2$ and $\gamma_f = 0.2\ \mathrm{m}^{-1}$. These values result in trajectories with speeds and maneuverability that are reasonable for micro UAVs. In particular, the maximum speed of each agent typically remains below 5.0 m/s. The agent orientation $\mathbf{q}_i^{(n)}$ evolves as follows: At each time step n, agent i rotates with random turn rate $\boldsymbol{\omega}_i^{(n)} \sim \mathcal{N}(\mathbf{0}, \sigma_\omega^2 \boldsymbol{I}_3)$, where $\sigma_\omega = 0.5\ \mathrm{s}^{-1}$. Note that $\boldsymbol{\omega}_i^{(n)}$ is the turn rate in the local reference frame of agent i. The corresponding state evolution model is provided in [30].

As in most inertial navigation techniques for multisensor fusion, in the simulated algorithm, the measurements provided by the IMU are incorporated as deterministic control input $\boldsymbol{u}_i^{(n)} = [\boldsymbol{u}_{i,\varphi}^{(n)\mathrm{T}} \boldsymbol{u}_{i,\omega}^{(n)\mathrm{T}}]^\mathrm{T}$. In particular, the IMU measurement $\boldsymbol{u}_i^{(n)}$ consists of an acceleration measurement $\boldsymbol{u}_{i,\varphi}^{(n)}$ and a turn-rate measurement $\boldsymbol{u}_{i,\omega}^{(n)}$, which are realizations of the RVs

$$\mathbf{u}_{i,\varphi}^{(n)} = \boldsymbol{\varphi}_i^{(n)} + \mathbf{c}_{i,\varphi}^{(n)}$$
$$\mathbf{u}_{i,\omega}^{(n)} = \boldsymbol{\omega}_i^{(n)} + \mathbf{c}_{i,\omega}^{(n)}\,,$$

where $\boldsymbol{\varphi}_i^{(n)}$ is the true acceleration of agent i in its local reference frame. The IMU noise $\mathbf{c}_i^{(n)} = [\mathbf{c}_{i,\varphi}^{(n)\mathrm{T}} \mathbf{c}_{i,\omega}^{(n)\mathrm{T}}]^\mathrm{T}$ consists of acceleration $\mathbf{c}_{i,\varphi}^{(n)} \sim \mathcal{N}(\mathbf{0}, \sigma_\varphi^2 \boldsymbol{I}_3)$ and turn rate $\mathbf{c}_{i,\omega}^{(n)} \sim \mathcal{N}(\mathbf{0}, \sigma_\omega^2 \boldsymbol{I}_3)$ components. The functional form of the resulting state-evolution model $\mathbf{x}_i^{(n)} = a_i(\mathbf{x}_i^{(n-1)}, \mathbf{c}_i^{(n)}; \boldsymbol{u}_i^{(n)})$ is provided in [30].

The range measurement $\mathrm{z}_{ij}^{(n)}$ made by agent $i \in \mathcal{N}_\mathrm{a}$ with node j at time step n is modeled as

$$\mathrm{z}_{ij}^{(n)} = \big\| \mathbf{p}_i^{(n)} - \mathbf{p}_j^{(n)} \big\| + \mathrm{v}_{ij}^{(n)},$$

where $\mathrm{v}_{ij}^{(n)} \sim \mathcal{N}(0, \sigma_\mathbf{v}^2)$ is the Gaussian noise with standard deviation $\sigma_\mathbf{v} = 0.1$ m. A more detailed, technology-specific model for ranging from wideband radio-frequency signals can be found in [27] and [28].

It is assumed that the number of available channels for performing range measurements is limited to 16, which means that only a subset of 16 agents can perform range measurements at a specific time step n. For this reason, time-division multiple access (TDMA) is performed by partitioning 512 agents into 32 disjoint groups, with each group consisting of 16 agents. At each time step n, only the agents in one of the 32 groups can make range measurements while all the others remain idle. At each time step n, for those 16 agents that perform range measurements, the set $\mathcal{A}_i^{(n)}$ is given as follows: range measurements can only be performed with nodes that are located within a communication range of 52 m. Moreover, if there are more than M potential-neighbor nodes, M

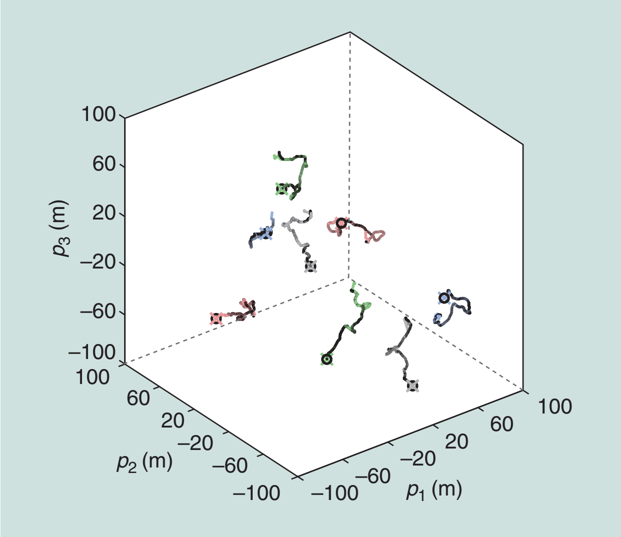


FIGURE 5. The trajectories related to eight exemplary agents and one simulation run. Colored and black curves represent the estimated and true trajectories, respectively. Similarly, colored crosses and black circles represent the estimated and true positions at the last time step, respectively.

of them are randomly selected. This selection of at most M neighbor nodes limits the energy consumption. It also reduces the number of loops in the factor graph in Figure 3 and thus the related negative effects, e.g., overconfident beliefs. Note that the communication range of 52 m was chosen so that for agents inside the region $\mathcal{R}$, there is at least one, and at most, four neighbor anchors.

In our simulation, the SPBP algorithm [5] is used, which is based on the design framework presented in the "Distributed Network-Localization Algorithms" section. Note that, in the considered scenario with 512 UAVs, localization algorithms based on SBE are unfeasible because they are not scalable in the number of agents [6, Sec. VII-C]. To the best of our knowledge, SPBP is the only available algorithm for cooperative location and orientation estimation in 3-D. One hundred simulation runs were performed and 1,200 time steps were simulated. Examples of true and estimated trajectories are shown in Figure 5. As a metric for localization performance the 3-D localization error outage (LEO) was used. The outage is a well-established concept in wireless communications; in the context of NLN, the LEO is similarly defined as the empirical probability that the localization error is above the predefined threshold e_{th}.

Network localization results

To study the impact of cooperation among agents as well as the impact related to multisensor fusion, the following configurations are compared: In the "Baseline" configuration, an agent makes range measurements only with the anchors within its communication range and does not perform IMU measurements. In the "Spatial Cooperation" configuration, additional range measurements are performed by cooperation among agents. In the "IMU Fusion" configuration, IMU measurements are performed but agents do not cooperate. Finally, in the "Spatial Cooperation + IMU Fusion" configuration, cooperation among agents as well as IMU measurements are performed. Note that for the network localization results presented in this section the following parameters were used: $M = 6, \sigma_{\varphi} = 10^{-4}\ \mathrm{m/s}^2$, and $\sigma_{\omega} = 5 \times 10^{-3}\mathrm{s}^{-1}$.

Figure 6 shows the LEOs (obtained by averaging more than 100 simulation runs, 512 agents, and 900 time steps) versus threshold e_{th} for the four simulated configurations. Since SPBP needs a certain number of time steps for initialization, for the network localization results, the first 300 time steps were not incorporated in the LEOs evaluation. The following key observations can be obtained from these results:

1) A very desirable localization performance can be obtained with "Spatial Cooperation + IMU Fusion." Specifically, the threshold e_{th} is 0.11 m and 0.17 m at a LEO of 10^{-1} and 10^{-2}, respectively. Remarkably, for $e_{\mathrm{th}} \geq 0.3$ m the LEO is 0.
2) The localization error is significantly reduced by spatial cooperation. In particular, by comparing "IMU Fusion" with "Spatial Cooperation + IMU Fusion," it can be seen that the e_{th} is reduced from 0.54 m to 0.11 m (by 79.6%) at a LEO of 10^{-1} and from 4.21 m to 0.17 m (by 96.0%) at a LEO of 10^{-2}. The reason for the performance gain of "Spatial Cooperation + IMU Fusion" with respect to "IMU Fusion" is that in the former configuration the agents have more neighbor nodes available for localization.
3) Incorporating IMU measurements also significantly reduces the localization error. More specifically, by comparing "Spatial Cooperation" with "Spatial Cooperation + IMU Fusion" it can be seen that the e_{th} is reduced from 2.92 m to 0.11 m (by 96.2%) at a LEO of 10^{-1} and from 5.00 m to 0.17 m (by 96.0%) at a LEO of 10^{-2}. The performance benefits "Spatial Cooperation + IMU Fusion" can be explained by the fact that the agents only makes range measurements every 32 time steps. Using "Spatial Cooperation" the localization error accumulates rapidly during the time period when no range measurements are performed. In contrast, by incorporating IMU measurements as in "Spatial Cooperation + IMU Fusion," this localization error accumulation can be significantly reduced.
4) Due to the few neighbor nodes available for localization and the high mobility of the agents, the localization performance of "Baseline" is very poor.

Network operation results

To demonstrate the benefits of network operation algorithms, a heterogeneous network that consists of two UAV classes was simulated. There were 256 UAVs in each class. The first class performed IMU measurements with noise standard deviations $\sigma_{\varphi} = 3.3 \times 10^{-5}\ \mathrm{m/s}^2$ and $\sigma_{\omega} = 1.7 \times 10^{-3}\mathrm{s}^{-1}$; the second class performed IMU measurements with noise standard deviations $\sigma_{\varphi} = 3 \times 10^{-4}\ \mathrm{m/s}^2$, and $\sigma_{\omega} = 1.5 \times 10^{-2}\mathrm{s}^{-1}$. All other parameters are as described in the "Scenario" section and were identical for both classes. For "Node Activation," spatial cooperation and IMU fusion was simulated together with the distributed network activation algorithm described in the "Distributed Node Activation" section to control the RMU measurements. At each time step n, every UAV determined its channel access probability $\zeta_i^{(n)}$ according to (28) and if $\zeta_i^{(n)} = 1$, it tried to access the channel. In a certain step, if multiple UAVs of the same group (see the "Case Study" section) that were also in the same subnetwork tried to access the channel, only one randomly selected UAV was able to perform an RMU measurement. As a reference method, "TDMA" was simulated where, as in the previous "Network Localization Results" section, spatial cooperation and IMU fusion with TDMA for channel access was performed. For both "Node Activation" and "TDMA," $M = 4$ and $M = 6$ were considered.

In the simulated scenario, "Node Activation" had a number of communication links related to RMU measurements that, compared to "TDMA," was reduced by 14.2% and 19.9% for $M = 4$ and $M = 6$, respectively. The average number of measurements performed per agent and per time step was 0.13 $(M = 4)$ and 0.19 $(M = 6)$ for "TDMA" and 0.11 $(M = 4)$ and 0.15 $(M = 6)$ for "Node Activation." Furthermore, consider a UWB radio that consumes 1.7×10^{-4} J per range measurement was used as an RMU [16], "Node Activation" can reduce the overall energy consumption of the network for all 1,200 time steps by 2.1 J for $M = 4$ and by 4.2 J for $M = 6$.

Figure 7 shows the LEOs—obtained by averaging more than 100 simulation runs, 512 agents, and 1,200 time steps—versus threshold e_{th} for the four simulated configurations. Note that the "Spatial Cooperation + IMU Fusion" results in Figure 6 correspond to the identical scenario as the "TDMA," $M = 6$ results in Figure 7. However, contrary to Figure 6, in Figure 7 all 1,200 time steps are considered. The following two observations can be made:

1) "Node Activation" can significantly increase localization accuracy. In particular, at a LEO of 10^{-2}, e_{th} is reduced from 7.18 m to 2.29 m, i.e., by 68.1% for $M = 4$ and from 6.42 m to 0.85 m, i.e., by 86.8% for $M = 6$. This is because with "Node Activation" UAVs in the second class tend to perform more range measurements compared to the ones in the first class, so they can compensate for their larger IMU noise standard deviation. In this way, "Node Activation" can also reduce the overall localization error of the network compared to "TDMA," where the UAVs in both classes make the same number of range measurements

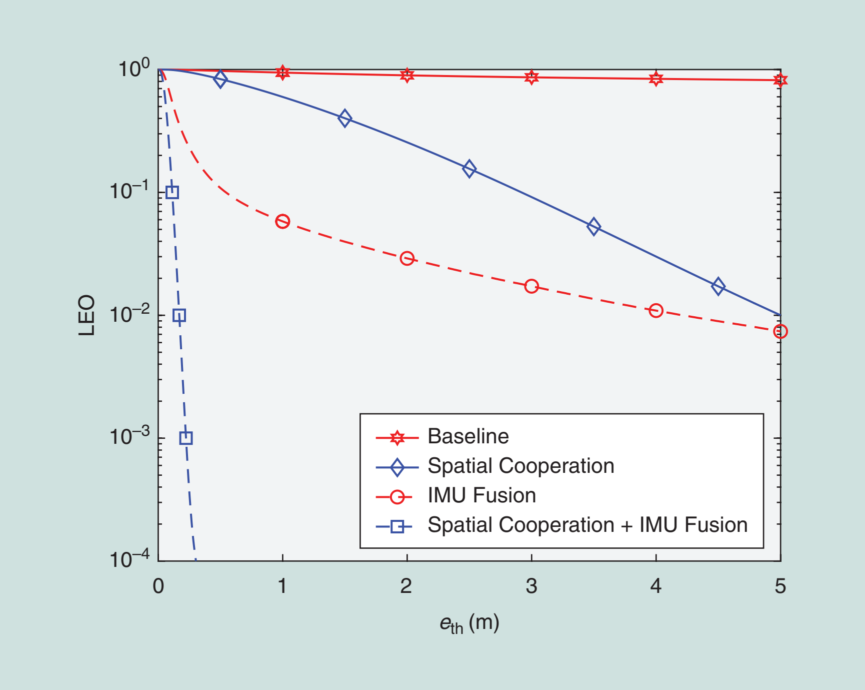


FIGURE 6. The LEO versus threshold e_{th} for the different simulated network localization configurations.

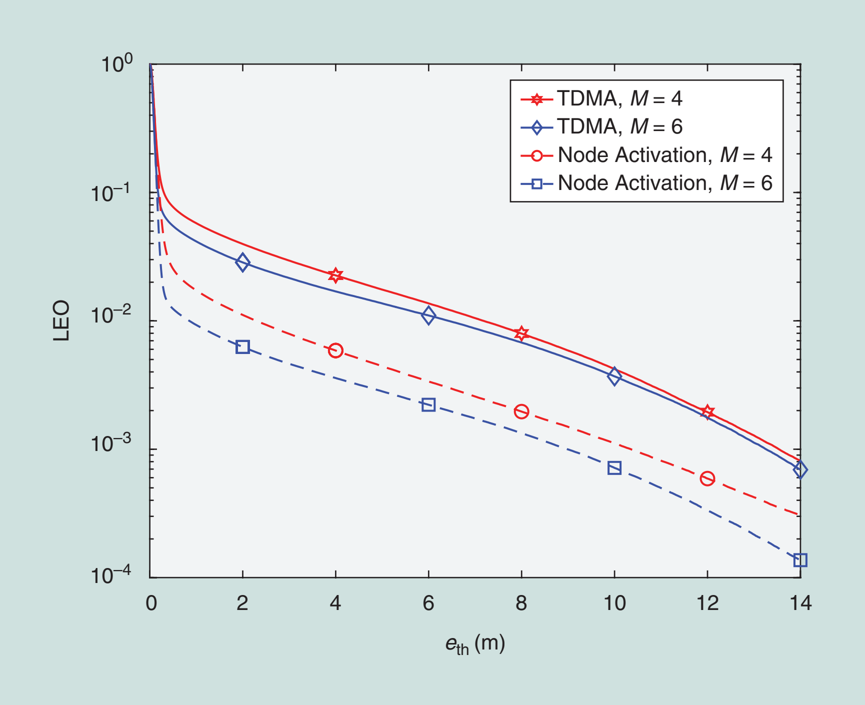


FIGURE 7. The LEO versus threshold e_{th} for the different channel access strategies and different M.

on average. Note that the improvement in localization performance is most significant at the first time steps during the initialization phase of the algorithm.

2) Incrementing M from four to six results is a localization error reduction that is small compared to the reduction related to performing "Node Activation" instead of "TDMA."

In particular, "Node Activation" for $M = 4$ performs significantly better than "TDMA" for $M = 6$. Thus it can be noted that a smart activation of agents can compensate for a low number of neighboring nodes.

Final remarks

The size and heterogeneity of IoT networks calls for a new class of scalable and technology-agnostic localization algorithms. In this article, we presented NLN, a paradigm that introduces scalable and distributed techniques for multisensor fusion in the IoT. NLN can provide technology-agnostic algorithms for IoT networks that exploit spatiotemporal cooperation to reduce the amount of required infrastructure. It also leads to the development of intelligent network operation strategies that allocate localization resources (e.g., transmission power and channel access opportunity) to extend the energy consumption of devices and to increase the localization accuracy. Localization performance and saving in terms of communication costs and energy consumption have been demonstrated in a case study with 500 mobile agents that aim to infer their location and their orientation in 3-D space. In particular, node activation significantly reduced energy consumption and, at the same time, increases the localization performance of the network. These results confirmed that in IoT applications localization and navigation performance can be strongly increased by multisensor fusion and cooperation among devices.

The size and heterogeneity of IoT networks call for a new class of scalable and technology-agnostic localization algorithms.

Acknowledgments

This research was supported, in part, by the Office of Naval Research under Grants N00014-16-1-2141 and N62909-18-1-2017, Defense University Research Instrumentation Program under Grant N00014-17-1-2379, the Austrian Science Fund under Grant J3886-N31, the Hong Kong Innovation and Technology Commission under Grant ITS/066/17FP, and the European Union's H2020 research and innovation programme Marie Skłodowska-Curie under Grant 703893. We also thank B. Teague and T. Wang for stimulating discussions.

Authors

Moe Z. Win (win@mit.edu) is a professor at the Massachusetts Institute of Technology (MIT), and the founding director of the Wireless Information and Network Sciences Laboratory. Prior to joining MIT, he was with AT&T Research Laboratories and NASA Jet Propulsion Laboratory. His current research interests include network localization and navigation, network-interference exploitation, and quantum-information science. He has been an IEEE Distinguished Lecturer and, currently, is serving on the SIAM Diversity Advisory Committee. He is an elected fellow of the American Association for the Advancement of Science and the Institution of Engineering and Technology. He was honored with two IEEE Technical Field Awards: the IEEE Kiyo Tomiyasu Award and the IEEE Eric E. Sumner Award. Other recognitions include the IEEE Communications Society Edwin H. Armstrong Achievement Award, the International Prize for Communications Cristoforo Colombo, and the Copernicus Fellowship and the Laurea Honoris Causa from the Università degli Studi di Ferrara. He is an ISI highly cited researcher. He is a Fellow of the IEEE.

Florian Meyer (fmeyer@mit.edu) received his Dipl.-Ing. and Ph.D. degrees in electrical engineering from Technische Universität Wien, in 2011 and 2015, respectively. He is currently a postdoctoral associate at the Wireless Information and Network Sciences Laboratory, Massachusetts Institute of Technology. In 2016, he was a research scientist with the NATO Centre for Maritime Research and Experimentation. His research interests include inference on graphs, cooperative navigation, multiobject tracking, and information-seeking control. He served as a technical program committee member of several IEEE conferences and was cochair of the IEEE Workshop on Advances in Network Localization and Navigation at the IEEE International Conference on Communications in 2018. He is an Erwin Schrödinger Fellow and a Member of the IEEE.

Zhenyu Liu (zhenyu.liu.us@ieee.org) received his B.Sc. and M.Sc. degrees in electronic engineering from Tsinghua University, in 2011 and 2014, respectively. He received academic excellence scholarships from Tsinghua University from 2008 to 2010. In 2014, he joined the Wireless Information and Network Sciences Laboratory at the Massachusetts Institute of Technology, where he is pursuing his Ph.D. degree in aeronautics and astronautics. His research interests include statistical inference and stochastic optimization, with application to communication and localization networks. He received the Best Paper Award at the IEEE Latin-American Conference on Communications in 2017. He has translated his research on network localization and navigation into practical demonstration systems for which he received the first prize at the IEEE Communications Society Student Competition in 2016.

Wenhan Dai (whdai@mit.edu) received his B.S. degrees in electronic engineering and mathematics from Tsinghua University, in 2011 and his M.S. degree in aeronautics and astronautics at the Massachusetts Institute of Technology, in 2014, where he is pursuing his Ph.D. degree in aeronautics and astronautics with the Wireless Information and Network Sciences Laboratory. His research interests include communication theory and stochastic optimization with application to wireless communication and network localization. He was honored by the Marconi Society with the Paul Baran

Young Scholar Award in 2017. He received the Marconi-Bioinformatic, Intelligent Systems and Educational Technology Best Paper Award from the IEEE International Conference on Ubiquitous Wireless Broadband in 2017, the Chinese Government Award for Outstanding Student Abroad in 2016, and first prize at the IEEE Communications Society Student Competition in 2016. He was recognized as an exemplary reviewer of *IEEE Communications Letters* in 2014. He is a Student Member of the IEEE.

Stefania Bartoletti (stefania.bartoletti@unife.it) received her laurea degree (summa cum laude) in electronics and telecommunications engineering and her Ph.D. degree in information engineering from the University of Ferrara, in 2011 and 2015, respectively. She is currently a Marie Skłodowska-Curie Global Fellow at the University of Ferrara, within the H2020 European Framework for a research project with the Massachusetts Institute of Technology. Her research interests include the theory and experimentation of wireless networks for passive localization and physical behavior analysis. She served as chair of the Technical Program Committee of the IEEE Workshop on Advances in Network Localization and Navigation at the IEEE International Conference on Communications in 2017 and 2018, respectively. She is a recipient of the 2016 Paul Baran Young Scholar Award of the Marconi Society. She is a Member of the IEEE.

Andrea Conti (a.conti@ieee.org) received his Ph.D. degree in electronic engineering and computer science from the University of Bologna, in 2001. He is currently an associate professor with the University of Ferrara. In the summer of 2001, he was with the Wireless Systems Research Department, AT&T Research Laboratories. He is a frequent visitor of the Wireless Information and Network Sciences Laboratory at the Massachusetts Institute of Technology, where he holds the research affiliate appointment. His research interests include theory and experimentation of wireless systems and networks, including network localization and distributed sensing. He is a recipient of the HTE Puskás Tivadar Medal and of the IEEE Communications Society's Stephen O. Rice Prize in the field of Communications Theory. He is a cofounder and elected secretary of the IEEE Quantum Communications and Information Technology Emerging Technical Subcommittee. He is an elected fellow of the Institution of Engineering and Technology and has been selected as an IEEE Distinguished Lecturer.

References

[1] M. Z. Win, A. Conti, S. Mazuelas, Y. Shen, W. M. Gifford, D. Dardari, and M. Chiani, "Network localization and navigation via cooperation," *IEEE Commun. Mag.*, vol. 49, no. 5, pp. 56–62, May 2011.

[2] A. T. Ihler, J. W. Fisher, III, R. L. Moses, and A. S. Willsky, "Nonparametric belief propagation for self-localization of sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 4, pp. 809–819, Apr. 2005.

[3] Y. Shen and M. Z. Win, "Fundamental limits of wideband localization—Part I: A general framework," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4956–4980, Oct. 2010.

[4] A. Conti, M. Guerra, D. Dardari, N. Decarli, and M. Z. Win, "Network experimentation for cooperative localization," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 2, pp. 467–475, Feb. 2012.

[5] F. Meyer, O. Hlinka, and F. Hlawatsch, "Sigma point belief propagation," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 145–149, Feb. 2014.

[6] F. Meyer, O. Hlinka, H. Wymeersch, E. Riegler, and F. Hlawatsch, "Distributed localization and tracking of mobile networks including noncooperative objects," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 1, pp. 57–71, Mar. 2016.

[7] F. Zabini and A. Conti, "Inhomogeneous Poisson sampling of finite-energy signals with uncertainties in $\mathbb{R}^d$," *IEEE Trans. Signal Process*, vol. 64, no. 18, pp. 4679–4694, Sept. 2016.

[8] A. Sandryhaila and J. M. F. Moura, "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 80–90, Sept. 2014.

[9] D. Dardari, A. Conti, C. Buratti, and R. Verdone, "Mathematical evaluation of environmental monitoring estimation error through energy-efficient wireless sensor networks," *IEEE Trans. Mobile Comput.*, vol. 6, no. 7, pp. 790–802, July 2007.

[10] R. Karlsson and F. Gustafsson, "The future of automotive localization algorithms: Available, reliable, and scalable localization: Anywhere and anytime," *IEEE Signal Process. Mag.*, vol. 34, no. 2, pp. 60–69, Mar. 2017.

[11] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. IEEE*, vol. 92, no. 3, pp. 401–422, Mar. 2004.

[12] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag, 2001.

[13] J. H. Kotecha and P. M. Djuric, "Gaussian sum particle filtering," *IEEE Trans. Signal Process.*, vol. 51, no. 10, pp. 2602–2612, Oct. 2003.

[14] Z. Liu, W. Dai, and M. Z. Win, "Mercury: An infrastructure-free system for network localization and navigation," *IEEE Trans. Mobile Comput.*, vol. 17, no. 5, pp. 1119–1133, May 2018.

[15] W. Dai, Y. Shen, and M. Z. Win, "A computational geometry framework for efficient network localization," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1317–1339, Feb. 2018.

[16] B. Teague, Z. Liu, F. Meyer, and M. Z. Win, "Peregrine: 3-D network localization and navigation," in *Proc. IEEE Latin-American Conf. Communications (LATINCOM)*, Nov. 2017.

[17] Y. Bar-Shalom, X.-R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation: Theory Algorithms and Software*. Hoboken, NJ: Wiley, 2004.

[18] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[19] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Englewood Cliffs, NJ: Prentice Hall, 2000.

[20] F. Daum and J. Huang, "Curse of dimensionality and particle filters," in *Proc. IEEE Aerospace Conf. (AeroConf)*, 2003, pp. 1979–1993.

[21] W. Dai, Y. Shen, and M. Z. Win, "Distributed power allocation for cooperative wireless network localization," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 1, pp. 28–40, Jan. 2015.

[22] W. Dai, Y. Shen, and M. Z. Win, "Energy-efficient network navigation algorithms," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 7, pp. 1418–1430, July 2015.

[23] H. Godrich, A. P. Petropulu, and H. V. Poor, "Power allocation strategies for target localization in distributed multiple-radar architectures," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3226–3240, July 2011.

[24] T. Wang, Y. Shen, A. Conti, and M. Z. Win, "Network navigation with scheduling: Error evolution," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7509–7534, Nov. 2017.

[25] X. Shen and P. K. Varshney, "Sensor selection based on generalized information gain for target tracking in large sensor networks," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 363–375, Jan. 2014.

[26] T. Wang, B. Teague, and M. Z. Win, "Distributed situation-aware scheduling algorithm for network navigation," in *Proc. IEEE Int. Conf. Ubiquitous Wireless Broadband*, Sept. 2017.

[27] D. Dardari, A. Conti, U. J. Ferner, A. Giorgetti, and M. Z. Win, "Ranging with ultrawide bandwidth signals in multipath environments," *Proc. IEEE*, vol. 97, no. 2, pp. 404–426, Feb. 2009.

[28] S. Bartoletti, W. Dai, A. Conti, and M. Z. Win, "A mathematical model for wideband ranging," *IEEE J. Sel. Topics Signal Process*, vol. 9, no. 2, pp. 216–228, Mar. 2015.

[29] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[30] J. D. Hol, F. Dijkstra, H. Luinge, and T. B. Schön, "Tightly coupled UWB/IMU pose estimation," in *Proc. IEEE Int. Conf. Ultra-Wideband (ICUWB)*, 2009, pp. 688–692.

Matthew Stamm, Paolo Bestagini, Lucio Marcenaro, and Patrizio Campisi

Forensic Camera Model Identification

Highlights from the IEEE Signal Processing Cup 2018 Student Competition

Determining the make and model of the camera that captured an image has been an important research area in information forensics for more than a decade [1]–[3] (see Figure 1). Information about which type of camera captured an image can be used to help determine or verify the origin of an image and can form an important piece of evidence in some scenarios, such as analyzing images involved in child exploitation investigations. While metadata may contain information about an image's source camera, metadata is both easy to falsify and frequently missing from an image. As a result, signal processing researchers have developed information forensic algorithms that can exploit traces intrinsically present in the image itself.

Camera model identification was selected as the topic for this year's IEEE Signal Processing Cup (SP Cup) competition. The SP Cup is a student competition in which undergraduate students form teams to work on real-life challenges. Each team should include one faculty member as an advisor, at most one graduate student as a mentor, and at least three but no more than ten undergraduate students. These teams participated in an open competition, and the top three were selected to present their work at the final stage of the competition at the 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), hosted in Calgary, Alberta, Canada, on 15 April.

In this article, we share an overview of the IEEE SP Cup experience, including competition setup, teams, technical approaches, and statistics.

Camera model identification

To understand how information forensic algorithms are able to determine which type of camera captured an image, it is important to first review how a digital camera captures an image. A digital camera's internal processing pipeline is composed of several different components, as shown in Figure 2. Light enters a camera through its lens, which focuses the light on the camera's optical sensor. Because a sensor can typically measure only one of the three primary colors of light at each pixel location on the sensor, an optical filter known as a color filter array (CFA) is placed in between the lens and the sensor. The CFA, which normally consists of a repeating 2 × 2 pixel pattern, allows only one color band of light to fall incident upon the sensor at each pixel location.

Information about which type of camera captured an image can be used to help determine or verify the origin of an image.

The resulting image produced by the sensor consists of three partially sampled color layers in which only one color value is recorded at each pixel location. Next, the remaining two color values at each pixel location are interpolated through a process known as *demosaicing*. After this, the image may be subject to internal processing, such as white balancing and Joint Photographic Experts Group (JPEG)

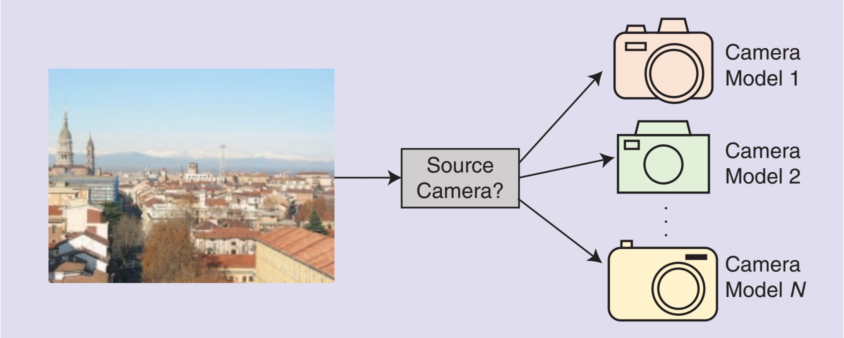


FIGURE 1. A representation of the camera model attribution problem: given an image, detect which camera model was used to shoot it within a closed set.

Digital Object Identifier 10.1109/MSP.2018.2847326
Date of publication: 28 August 2018

1053-5888/18©2018IEEE

compression, before the final output image is produced.

A significant amount of information forensics research has shown that many of the components that make up a camera's internal processing pipeline introduce statistical traces or other artifacts into an image. Because different camera models use different implementations of each physical and algorithmic component in their internal processing pipeline, the traces left in an image by each component can be linked to the make and model of the camera that captured the image.

Different camera models, for example, typically use different proprietary demosaicing algorithms to perform color interpolation. Several forensic algorithms have been developed to model and estimate the demosaicing filter used by a camera or to capture pixel value dependencies introduced by the demosaicing process [4]–[6]. Statistical models of sensor noise and other noise sources have been used to determine the make and model of an image's source camera [7], [8] as have traces left by proprietary quantization tables used by a camera during JPEG compression [9]. Additionally, statistical techniques from steganalysis [10] and heuristically designed feature sets [11] have been designed to capture camera model traces. Methods based on convolutional neural networks (CNNs) have also lately been proposed [12]–[14]. Several survey papers exist that can provide participating teams an overview of existing research in this area [1]–[3].

Broadly speaking, camera model identification algorithms typically operate by designing a signal processing algorithm to extract a particular forensic trace from an image. Camera model fingerprints are then learned by extracting traces from many images taken by a particular camera model and then repeating this process for several different camera models. After this, these traces are used as classification features when training a machine-learning algorithm, such as a support vector machine or neural network, to recognize an image's source camera model.

Tasks in the SP Cup 2018

The goal of this competition was for teams to build a system capable of determining the type of camera (manufacturer and model) that captured a digital image without relying on metadata. Teams were asked to use their signal processing expertise to extract traces from images that could be linked with different camera models. The competition was split into two stages, an open competition that any eligible team could participate in and an invitation-only final competition. The three teams reporting the highest scores in the open competition were invited to the final competition.

The goal of this competition was for teams to build a system capable of determining the type of camera that captured a digital image without relying on metadata.

The open competition part one: Unaltered images

Part one of the open competition was designed to give teams a well-defined problem and data set that they could use to become familiar with forensic camera model identification. Participants were provided with a data set that they could use to train and test their camera model identification systems. This data set consisted of images from ten different camera models (including point-and-shoot cameras, cell phone cameras, and digital single-lens reflex cameras), with 200 images captured using each camera model. All images were captured and stored as JPEGs using the device's default settings.

On the Kaggle website, a new evaluation data set was released. This data set contained unseen images that may have been captured using any of the ten camera models in the original data set (but taken by different devices). Teams were asked to use their systems to identify the camera model used to capture each image in the evaluation data set. The accuracy of each system was used as the score for each team.

To prevent brute force attempts to guess the camera model associated with each evaluation set image, teams were allowed to submit a limited number of classification attempts per day during the evaluation period. Additionally, images in the evaluation set were taken using different devices (but the same model) than those used to create the training data set. This prevented matches on the basis of an individual cameras sensor using photo response nonuniformity sensor noise traces as opposed to general traces left by all cameras of a common make and model.

The open competition part two: Edited images

Part two of the open competition was designed to present teams with a more challenging scenario: determining the source camera model of images that had been postprocessed. In this part of

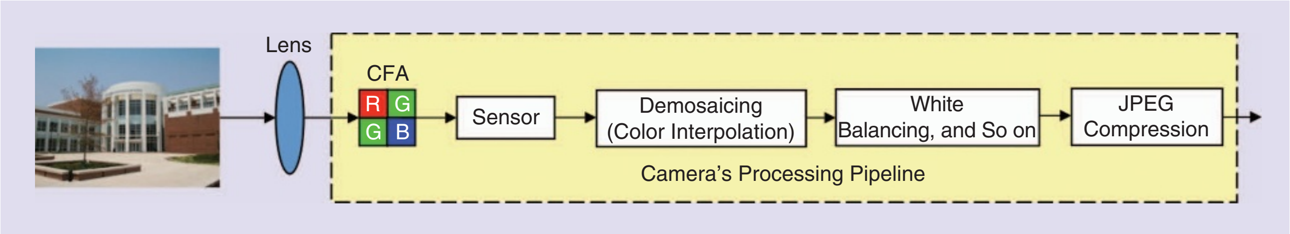

Lens
CFA
R
G
G
B
Sensor
Demosaicing (Color Interpolation)
White Balancing, and So on
JPEG Compression
Camera's Processing Pipeline

FIGURE 2. The camera acquisition processing pipeline. Light rays are focused by a lens to hit the sensor after passing through a CFA. Images acquired by the sensor are typically further processed and compressed to be stored into memory.

Winners of the SP Cup 2018

First Place: Team FIIGO

- University Federico II of Naples, Italy
- Undergraduate students: Gioacchino Gargiulo, Raffaele Mazza, and Fabrizio Guillaro
- Tutor: Davide Cozzolino
- Supervisor: Luisa Verdoliva
- Technical approach: Team FIIGO (Figure S1) proposed an approach based on an ensemble of state-of-the-art deep neural networks. Specifically, the proposed technique involved: 1) the fusion of a large number of network models, 2) the use of multiple patch sizes (i.e., networks were trained on small patches and fine-tuned on larger ones), 3) a maximum likelihood decision method to cope with small patches extracted densely from test images, and 4) identification of problematic classes and design of dedicated detectors. The proposed training strategy also benefitted from a great number of images downloaded on purpose to enrich the training data set.

Second Place: Team GPU_muscle

- National Research University Higher School of Economics, Moscow Institute of Physics and Technology, Russia
- Undergraduate students: Artur Fattakhov, Ilya Kibardin, and Andrey Kiselev
- Tutor: Artur Kuzin
- Supervisor: Valeriy Babushkin
- Technical approach: Team GPU_muscle (Figure S2) proposed an approach based on a convolutional neural network (CNN) ensemble. In particular, they started from eight pretrained CNN models and fine-tuned them on an enriched training set. This training set was acquired by downloading more than 500 GB of images from the web and filtering them based on model, resolution, and Joint Photographic Experts Group compression quality. The classification layer of each CNN was modified to accommodate for an additional binary input indicating the image being manipulated or not. The final ensemble was carried using geometric mean.

Third Place: Team blzr

- University of Alabama, Birmingham
- Undergraduate students: Sai Chintha, Ruta Bhat, and Mark Salazar
- Tutor: David Odaibo
- Supervisor: Leon Jololian

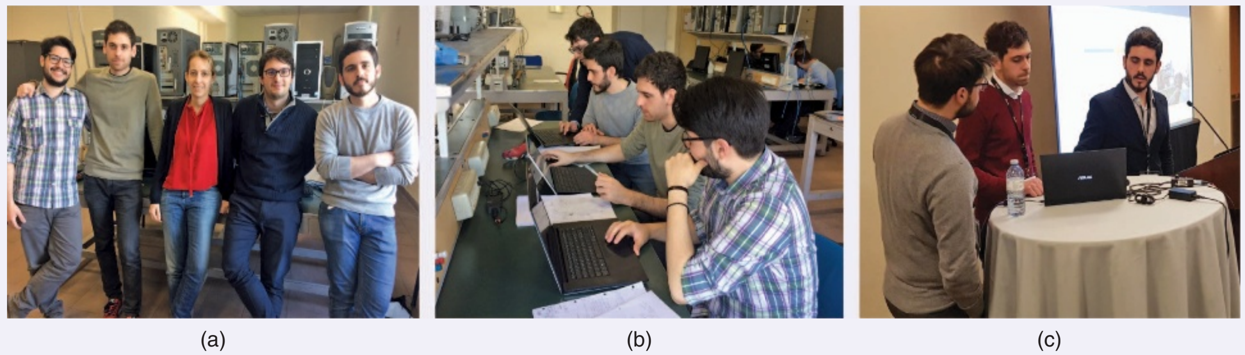

(a)
(b)
(c)

FIGURE S1. The first-place team: Team FIIGO. (a) The team members (from left): Raffaele Mazza, Fabrizio Guillaro, Luisa Verdoliva, Davide Cozzolino, and Gioacchino Gargiulo; (b) the team at work; and (c) the team at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

the competition, teams were presented with images that had been postprocessed using one of several operations, and they were asked to determine the make and model of the camera that captured the images. While postprocessing operations are commonly applied to images before they are shared online, these operations can potentially alter forensic traces present in the images.

In this competition, images were altered using postprocessing operations, such as JPEG recompression, cropping, and contrast enhancement. Teams were provided with a list of all possible postprocessing operations that were considered at the launch of the open competition. Additionally,

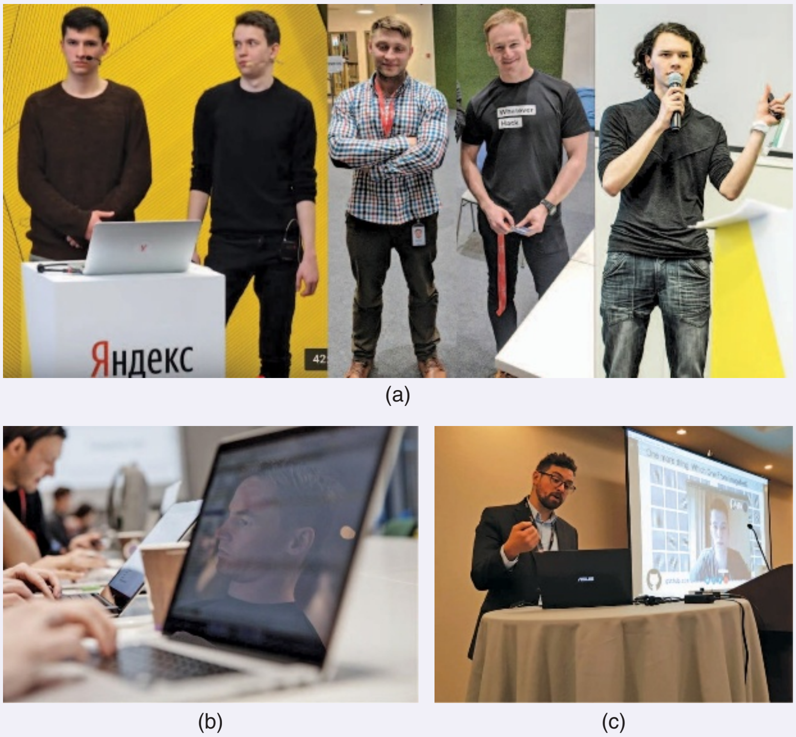


FIGURE S2. The second-place team: Team GPU_muscle. (a) The team members, (b) the team at work, and (c) Matthew Stamm introducing the team.

- Technical approach: Team blzr (Figure S3) proposed an approach based on a dual-input CNN. They used a densely connected CNN taking as input a 512 × 512 image crop and a single binary value indicating if the input image was altered or pristine. Specifically, the binary input was concatenated with the last layer of the base CNN. This concatenated vector was fed to a global average pooling layer followed by two additional densely connected layers with 512 and 128 hidden units and by softmax. The output of the network was used as the classification result. For network training, the team also synthetically performed data augmentation. Specifically, a set of editing operations was applied to all images in the data set.

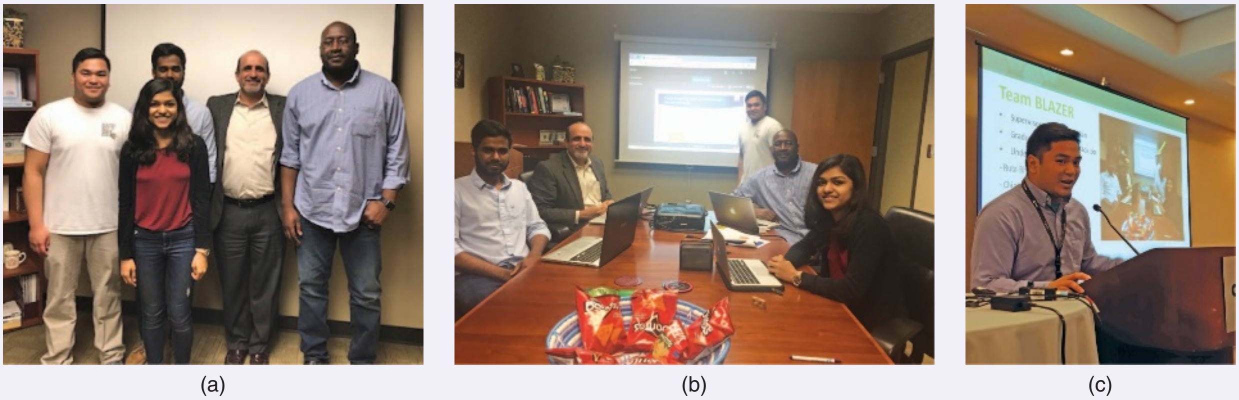


FIGURE S3. The third-place team: Team blzr. (a) The team members, (b) the team in the lab, and (c) the team at ICASSP.

teams were provided with a MATLAB script that they could use to generate postprocessed images from the original data set of unaltered images (this was intended to reduce the amount of data that each team had to download) and for training and testing purposes.

Similar to part one of the open competition, this portion of the competition was also run using Kaggle. The evaluation data set of postprocessed images came from the ten camera models used to create the original data set (but with different devices). Teams were asked to use their camera model identification system to determine which camera model captured each of these images.

The accuracy of each system was used as the score for each team. The final score of the open competition was a weighted average between part one (70%) and part two (30%) scores.

The open competition part three: Data set collection

This task is designed to give teams experience performing data collection. Teams were asked to capture 250 images using a camera model that was not provided in the original data set. Data collection guidelines were provided to teams along with data set upload instructions.

This portion of the competition did not contribute to the team's overall score. However, teams that did not participate in this task were considered ineligible to compete in the final competition. Data collected as part of this task were used to create an additional data set that will be released to the public to promote research efforts.

Final competition

The three highest-scoring teams from the open competition stage were selected as finalists and invited to compete in the final competition. These teams were provided with an additional data set consisting of images captured using new camera models. This data set contained both unaltered images and post-processed images. Teams were asked to update their systems and identify the camera models used to capture each image in the new data set.

Competition data

The data set, MATLAB scripts, and other associated material are available for download at http://misl.ece.drexel.edu/spcup2018/ as well as https://ieee-dataport.org/open-access/ieee-signal-processing-cup-2018-database-forensic-camera-model-identification.

Highlights of technical approaches

Due to the classification nature of the proposed task, all teams made use of supervised classification techniques borrowed from the machine-learning community. Teams were roughly split into two groups: 1) those making use of handcrafted features for camera model identification paired with a supervised classifier and 2) those exploiting completely data-driven approaches based on CNNs.

Despite some good results achieved with handcrafted approaches, the best results were obtained using CNNs. Among these techniques, a common pipeline was to feed multiple CNNs with images in the pixel domain (optionally cropped to smaller size) and join features obtained from multiple CNNs in an ensemble fashion. As teams were given both manipulated and unedited images, some groups proposed separate approaches for the two tasks. Other teams decided to embed into the classification layer a flag indicating whether the input image was edited or not.

Despite some good results achieved with handcrafted approaches, the best results were obtained using CNNs.

It is worth noting that the vast majority of the teams with higher accuracy benefit from an enriched training data set. Data-driven solutions trained on huge data sets of images downloaded from media-sharing platforms proved to achieve superior accuracy.

SP Cup 2018 statistics

The camera model identification competition was hosted on the Kaggle website (https://www.kaggle.com/c/sp-society-camera-model-identification/). In addition to student teams, a great number of additional teams (noneligible in terms of the SP Cup) also participated in the competition, for a total of 582 teams. Among these, 30 eligible student teams registered for the SP Cup. These teams came from many different countries all over the world: Europe (Italy, Switzerland, Germany, France, Spain, Greece), Asia (Russia, China, India, Bangladesh, and Iran), North America (United States), and Africa (Egypt).

As in previous years, the SP Cup was run as an online class through the Piazza platform, which allowed a continuous interaction with teams. In total, 239 students registered for the course, and the number of contributions has been higher than 180. An archive of the class is available at https://piazza.com/ieee_sps/other/spcup2018/home.

Since its inception, the SP Cup has received generous support from MathWorks, the maker of the popular MATLAB and Simulink platforms. MathWorks kindly provided funding support to the SP Cup and free MATLAB licenses to competing students.

Participants' feedback

In this section, we provide some feedback and perspectives received from the three winning teams.

Team FIIGO

- "I loved the idea of cooperating with a team and increasing my knowledge of deep learning. I gained a deeper insight into how neural networks really work and how to approach problems like this. It has been a great way to get some experience on this matter."

  —Fabrizio Guillaro, undergraduate

- "I liked the thrill of competing against teams coming from all over the world. This was also an opportunity to learn more about CNNs, camera model identification, and, more in general, digital image forensics. I learned how to use CNNs to work with images. I also learned that by carefully analyzing the data set and the problem, and by testing innovative solutions, you can achieve good results even without powerful and expensive hardware."

  —Raffaele Mazza, undergraduate

- "I loved competing against teams from all over the world. I gained a deeper understanding of CNNs and teamwork mechanics, learning how to coordinate efforts in a complex project."

  —Gioacchino Gargiulo, undergraduate

- "It was a nice experience. I liked working in the FIIGO team. We were all very motivated to give the best because the submission system provided the ranking in real time, and there were many strong teams. Thanks to the SP Cup competition, I have increased my experience on machine learning and deep learning. I learned new lessons on the importance of data augmentation and ensemble methods for classification."

  —Davide Cozzolino, tutor

- "It was a very stimulating competition. It was also great to work every day with the students and help them facing a real problem as part of a team. I learned the importance of using a good and large training set to obtain competitive results. Also, I had first-hand confirmation that the use of very deep networks is essential to achieve robustness."

  —Luisa Verdoliva, supervisor

Team GPU_muscle

- "I liked the atmosphere of challenge and the nonstandard task that was provided to solve. I also like that it is truly international. Personally I learned a new framework for deep learning and have applied techniques I have read before, such as snapshot ensembling and test-time augmentation. I also learned to work under the strict condition of a competition deadline."

  —Valeriy Babushkin, supervisor

- "The task itself was quite unusual and looked like a nice playground to test unconventional methods. Once again I was astonished by an unreasonable effectiveness of conventional CNNs. Results achieved with standard neural networks are quite impressive, despite the unusual task."

  —Andrey Kiselev, undergraduate

- "I learned to use standard but very effective tricks, and unexpectedly neural networks worked best in this task."

  —Artur Fattakhov, undergraduate

- "An interesting task, where deep convolution networks showed outstanding performance, although they were not developed for this type of task. I learned about team management and using PyTorch in a multi-graphics processing-unit scenario."

  —Artur Fattakhov, undergraduate

- "As did the other members of our team, I liked that the task was extraordinary and therefore interesting. Also, I liked that the competition was quite challenging because we had to compete with scientists with publications related to the task. I learned how to work with large data sets (we downloaded 500+ GB of external data during the first stage) and how to efficiently train neural networks for classification using such techniques as cyclic learning rate and snapshot ensembling. In addition, I learned not to trust local validation too much because before we downloaded external data, models were overfitting hard to features of individual devices."

  —Ilya Kibardin, undergraduate

Team blzr

- "I liked the inclusion of the undergraduates as part of the team. I think it provided all of us with an opportunity and our first foray into machine learning/artificial intelligence and competitive data science. I learned a lot about general neural networks and the specialized ones like densely connected convolutional networks and CNNs. I learned about fine-tuning certain parameters, data transformations that can help these general models to learn the subtleties involved in image forensic analysis."

  —Sai Chintha, undergraduate

- "The SP Cup gave me a broader perspective on the data analysis part and how it can be applied to real-world problems. I learned about neural networks, CNNs, different libraries like Keras that can be used for machine learning, the importance of training, cross-validation and testing data, data augmentation techniques, and different ways to go about solving the same problem."

  —Ruta Bhat, undergraduate

- "The Cup gave us an opportunity to work on an interesting problem. I learned about CNNs, training CNNs, and the importance of data augmentation to improve model performance."

  —Mark Salazar, undergraduate

- "It was an interesting challenge that gave us an opportunity to explore the limits of AI in imaging forensics. I learned that various neural network training strategies and densely connected CNNs are really good at image forensics."

  —David Odaibo, tutor

- "The SP Cup competition gave our undergraduate student team members an opportunity to learn and be exposed to machine learning and deep learning. Everyone on the team gained added appreciation for the technology and its applications. Next year, we will encourage more students to participate in such an educational and instructive competition. We were able to explore CNNs and get an added appreciation for how they are really good in applying to image analysis and, in particular, digital forensics. Our undergraduate students in the team thoroughly enjoyed learning about the inner workings of the algorithms and have renewed interest in pursuing graduate studies in the field of machine learning."

  —Leon Jololian, supervisor

Forthcoming project competitions for undergraduates

The sixth edition of the SP Cup will be held at ICASSP 2019. The theme of the 2019 competition will be announced in September. Teams who are interested in the SP Cup competition may visit this link: https://signalprocessingsociety.org/get-involved/signal-processing-cup.

In addition to the SP Cup, the IEEE Signal Processing Society (SPS) recently launched the Video and Image Processing (VIP) Cup. The second edition of the VIP Cup will be held at the IEEE International Conference on Image Processing in Athens, Greece, 7–10 October. The theme of this competition is "Lung Cancer Radiomics—Tumor Region Segmentation." For details, visit: https://signalprocessingsociety.org/get-involved/video-image-processing-cup.

Teams were asked to use their systems to identify the camera model used to capture each image in the evaluation data set.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under grant 1553610. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

As the SP Cup 2018 Organizing Committee, we would like to express our gratitude to all of the people who made this adventure a reality: the participating teams, the judging panel, the local organizers, the IEEE SPS Membership Board for its full support, Kaggle for letting us run the SP Cup through their system, and MathWorks for its sponsorship. A special thank you to Patrizio Campisi for his unmatched support throughout the whole competition.

Authors

Matthew Stamm (mstamm@drexel.edu) received his B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Maryland, College Park. He is an assistant professor with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia. He serves as a member of the IEEE Information Forensics and Security Technical Committee (IFS TC) and the editorial board of the IEEE Signal Processing Society repository. He initiated this edition of the Signal Processing Cup endorsed by the IEEE IFS TC. He is a Member of the IEEE.

Paolo Bestagini (paolo.bestagini@polimi.it) received his B.Sc. and M.Sc. degrees in telecommunications engineering and his Ph.D. degree in information technology from the Politecnico di Milano, Italy. He is an assistant professor at the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy. He is an elected member of the IEEE Information Forensics and Security Technical Committee. He ran the competition alongside Matthew Stamm. He is a Member of the IEEE.

Lucio Marcenaro (lucio.marcenaro@unige.it) received his M.Sc. degree in electronic engineering and his Ph.D. degree in computer science and electronic engineering from the University of Genova, Italy. He is an assistant professor with the Department of Electrical, Electronics and Telecommunication Engineering and Naval Architecture, University of Genoa, Italy. He chairs the Student Service Committee of the IEEE Signal Processing Society supporting the Signal Processing Cup. He is a Member of the IEEE.

Patrizio Campisi (patrizio.campisi@uniroma3.it) received his M.Sc. degree in electronic engineering from the Università degli Studi di Roma "La Sapienza," Italy, and his Ph.D. degree in electrical engineering from the Università degli Studi "Roma TRE," Italy. He is a professor at Roma Tre University, Rome, Italy. He is the former chair of the Student Service Committee of the IEEE Signal Processing Society and the current chair of the IEEE Information Forensics and Security Technical Committee. He has been the organizer of the Signal Processing Cup since 2015 and is the initiator of the Video and Image Processing Cup. He is a Senior Member of the IEEE.

References

[1] M. C. Stamm, M. Wu, and K. J. R. Liu, "Information forensics: An overview of the first decade," *IEEE Access*, vol. 1, pp. 167–200, May 2013.

[2] A. Piva. (2013). An overview on image forensics. *ISRN Signal Process.* [Online]. Available: https://www.hindawi.com/journals/isrn/2013/496701/

[3] M. Kirchner and T. Gloe, "Forensic camera model identification," in *Handbook of Digital Forensics of Multimedia Data and Devices.* Hoboken, NJ: Wiley, 2015, pp. 329–374.

[4] A. Swaminathan, M. Wu, and K. J. R. Liu, "Digital image forensics via intrinsic fingerprints," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 1, pp. 101–117, Mar. 2008.

[5] H. Cao and A. C. Kot, "Accurate detection of demosaicing regularity for digital image forensics," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 4, pp. 899–910, Dec. 2009.

[6] C. Chen and M. C. Stamm, "Camera model identification framework using an ensemble of demosaicing features," in *Proc. IEEE Int. Workshop Information Forensics and Security*, 2015, Rome, Italy, pp. 1–6.

[7] T. Filler, J. Fridrich, and M. Goljan, "Using sensor pattern noise for camera model identification," in *Proc. IEEE Int. Conf. Image Processing*, 2008, San Diego, CA, pp. 1296–1299.

[8] T. H. Thai, R. Cogranne, and F. Retraint, "Camera model identification based on the heteroscedastic noise model," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 250–263, Jan. 2014.

[9] E. Kee, M. K. Johnson, and H. Farid, "Digital image authentication from JPEG headers," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 3, pp. 1066–1075, Sept. 2011.

[10] F. Marra, G. Poggi, C. Sansone, and L. Verdoliva, "Evaluation of residual-based local features for camera model identification," in *Proc. Int. Conf. Image Analysis and Processing*, 2015, pp. 11–18.

[11] M. Kharrazi, H. T. Sencar, and N. Memon, "Blind source camera identification," in *Proc. Int. Conf. Image Processing (ICIP)*, 2004, pp. 709–712.

[12] A. Tuama, F. Comby, and M. Chaumont, "Camera model identification with the use of deep convolutional neural networks," in *Proc. IEEE Int. Workshop Information Forensics and Security (WIFS)*, 2016. doi: 10.1109/WIFS.2016.7823908.

[13] L. Bondi, L. Baroffio, D. Güera, P. Bestagini, E. J. Delp, and S. Tubaro, "First steps toward camera model identification with convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 259–263, Mar. 2017.

[14] D. Güera, S. K. Yarlagadda, P. Bestagini, F. Zhu, S. Tubaro, and E. J. Delp, "Reliability map estimation for CNN-based camera model attribution," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, 2018, 964–973. doi: 10.1109/WACV.2018.00111.

Pragnan Chakravorty

What Is a Signal?

After decades of advances in signal processing, this article goes back to square one, when the word *signal* was defined. Here we investigate if everything is all right with this stepping stone of defining a signal.

Relevance

We may remind ourselves of our high school days when not using the term *rectilinear motion*, i.e., motion in a straight line [1], in the definition of Newton's first law, would be grossly unacceptable to our instructors and the scientific community at large. As a matter of fact, the same accuracy is expected while framing any scientific statement. Unfortunately, there may be few exceptions to this practice of precision, and perhaps such an exception occurs when we try to answer the question "What is a signal?"

Although there is a common agreement that a signal in its most generic sense is rather abstract, it should not be the case with engineers particularly dedicated to signal processing. Moreover, our acceptance of a somewhat unclear definition of any subject challenges our basic understanding of it; this makes our investigation all the more relevant.

Prerequisites

As we go ahead with our task, general knowledge of algebra, functions of multiple variables, trigonometry with compound angles, and angular frequency are expected. In short, knowledge of undergraduate electrical or electronics engineering shall be more than a sufficient prerequisite.

Problem statement and solution

We pull some statements from a few of the established references that define a signal. Reference [2] states, "The signals, which are functions of one or more independent variables, contain information about the behavior or nature of some phenomenon." As to [3], "A signal, as the term implies, is a set of information or data." In [4], a signal is formally defined as "a function of one or more variables that conveys information on the nature of a physical phenomenon."

The unquantifiable

The unquantifiable features of the words or phrases, like *behavior, nature, some phenomenon,* and so on, stand to reason as to why the definition of *signal* is vague. Conversely, the word *information* is concrete and has been defined in information theory [5], [10]. Therefore, the use of this word in the definition of a signal should be consistent with its definition that is already in use. A quick look in to this statement from [6]: "the amount of information received from a message is directly related to the uncertainty or inversely related to the probability of its occurrence" reveals that the entities that we had been classifying as deterministic signals all this while (sinusoids, step, signum, and so on), don't fit into the definition of a signal as they don't contain any information.

Not everything is a signal

Further reading into [2]–[4] tells that in common practice, unlike what is traditionally defined, a signal is anything that is a function of time, space, or both. We may unanimously agree that something that varies as a function of time is a signal; however, if that is a function of only space, then practically anything that we sense or see is a signal. Obviously, none of us would want to define a signal as *anything* or *everything.*

It is now quite apparent that the definition of a signal, in the current literature, doesn't bring sufficient clarity. Therefore, the problem to be solved is to accurately define the meaning of the term signal, considering its wide scope and usage.

Observable change

Normally, our common understanding would not allow static things around us, which are functions of only space, to be called *signals*; indeed, we want something dynamic for that. However, an image, a function of only space, is believed to be a signal as evident from this statement from [4]: "An image is an example of a two-dimensional signal, with the horizontal and vertical coordinates of the image representing the two dimensions." Intuitively, we

Digital Object Identifier 10.1109/MSP.2018.2832195
Date of publication: 28 August 2018

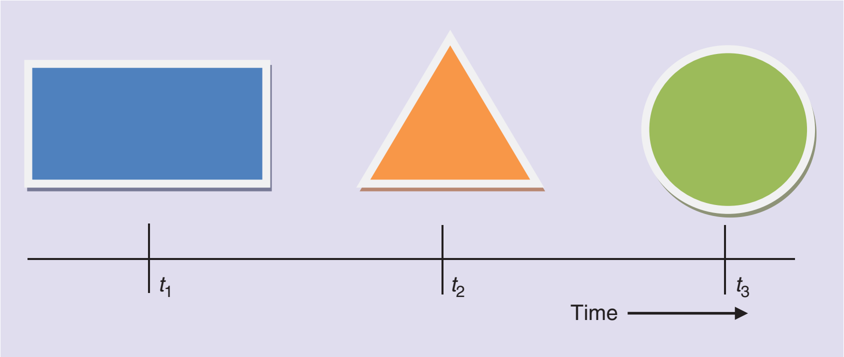


FIGURE 1. An image, with different geometrical shapes, shows changes only if each geometry is seen at a different time (i.e., t_1, t_2, and t_3); however, when the geometries are observed together at the same instant of time, no changes can be seen.

may agree that a signal must contain a change that can be observed. Now the question is, "Does an image convey any observable change?" Yes, it does. Consider the image in Figure 1. As we go from left to right, we do observe lot of changes; nevertheless, these changes could only be sensed because the observations were made at different times. Had we seen the entire image at the same instant, it would not have conveyed any observable change. Therefore, to convey an observable change, an entity must vary with time. So, an image, three dimensional (3-D) or two dimensional (2-D), may not be a signal itself but may serve as a storage for signals, much like digital storage oscilloscopes or graphs, when one of its dimensions is essentially a representation of time.

Lumped and distributed

At this point, it is important to know what exactly an image, or a similar entity, is. It is stated in [3], "When an electrical charge is distributed over a body, for instance, the signal is the charge density, a function of space rather than time." Notwithstanding the idea of a signal, such a charge distribution is simply known as *distributed charge* in its core literature [8], [9]; similarly, with respect to their distribution in space, entities like charge, mass, voltage, and so on, can either be lumped/point (not a function of space) or distributed (function of space); the distribution over space may be one dimensional (1-D) to 3-D. Electrical elements like capacitance, inductance, and resistance may also be either lumped or distributed [8]. Precisely, an amplitude is a direct measure of such lumped or distributed entities. Therefore, a distributed charge and a point charge may generally be termed as a *distributed amplitude* and a *point amplitude*, respectively. We may now try to convince ourselves that an image without a time dimension is a distributed amplitude of color(s). Practically, an image is a distribution of colored dots over space where a color is a combination of at least three color amplitudes; i.e., red, green and blue. Each of these dots may be viewed as a point amplitude or source. Additionally, a signal generated from a point amplitude may then be called a *point source signal* (*PSS*) and the one from a distributed amplitude may be called a *distributed source signal* (*DSS*). Since any quantifiable entity may be represented as an image, we show a unified illustration in Figure 2. An image of a mallet [Figure 2(a)] is shown, decreasing in size over time making an "image-signal" or a motion picture/movie as shown

After observing (1)–(3), we may be tempted to construe that a wave is nothing but a DSS. Although every wave is a DSS, every DSS may not be a wave.

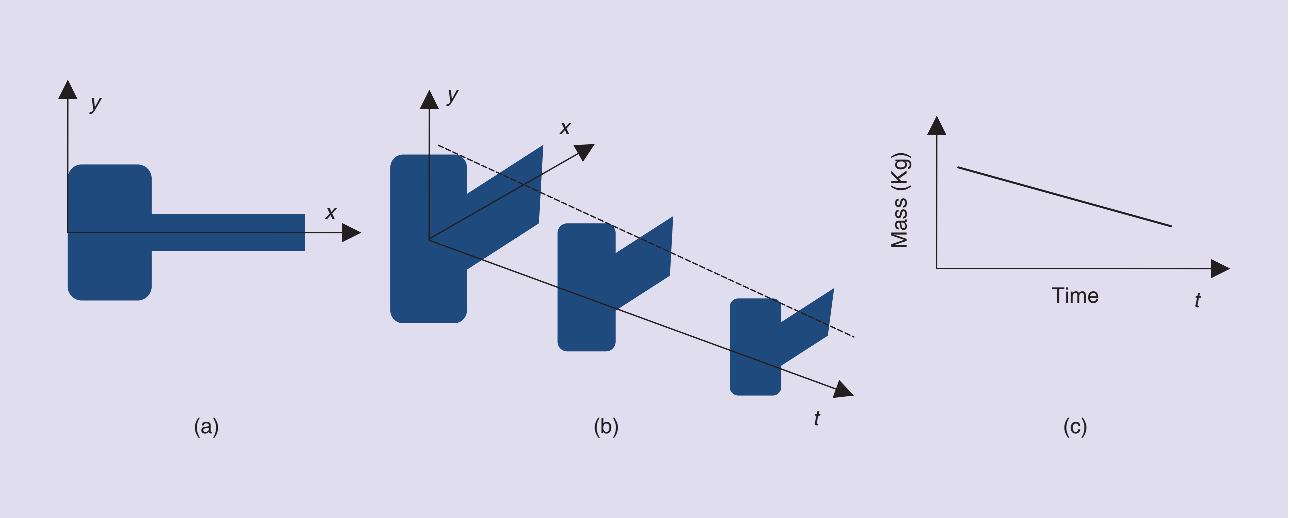


FIGURE 2. (a) A mallet shown as a distributed amplitude of mass. (b) A time-varying image of a mallet as 2-D distributed signal. (c) The overall mass as a lumped amplitude plotted against time to make it a lumped or point source signal.

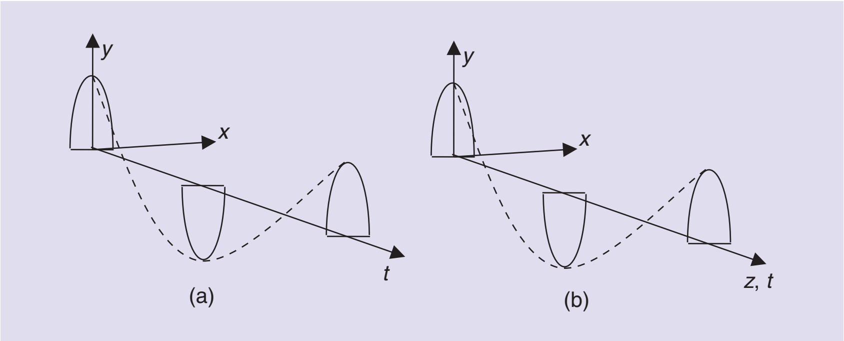


FIGURE 3. (a) and (b) are rough sketches of (4) and (5), respectively.

in Figure 2(b). The image of the mallet may also be seen as a distributed mass. Here, as the image decreases in size over time, so does the mass of the mallet. Interestingly, an image in form of a time graph, i.e., with time as one of its dimension, despite being distributed in nature, can also represent or store a PSS. For example, the change in the overall mass with respect to time can be seen as a PSS plotted in a time graph of Figure 2(c). Two-dimensional and 3-D distributed amplitudes in form of images are also prevalent in spatial signal processing done over beamforming antenna arrays [7]. Certainly, when these amplitudes are time varying, they then become 2-D or 3-D DSSs. Fortunately, though an image may not be a signal, processing it may unquestionably be called *signal processing* because images can't be processed without a change in time.

Signals and waves

Another common, and somewhat flawed, practice is to interchangeably use the words *signal* and *wave*. For example, it's very common to say or write the phrases *square wave, triangular wave*, and so on, without caring about the fact that waves are a special case of signals. The disparity between waves and other signals can be easily known by their mathematical expressions. Equation (1) is a signal with voltage as point amplitude; (2) is a nondispersive, 1-D traveling wave with wave number k [8], [9]; and (3) is a 1-D standing wave.

$$V(t) = V_o \sin(wt) \quad (1)$$

$$V(x,t) = V_o \sin(wt\text{-}kx) \quad (2)$$

$$V(x,t) = V_o \sin(wt)\sin(kx). \quad (3)$$

The symbols V/V_o, w, and t are voltage, angular frequency, and time, respectively [2]–[4]. The role of k with space x is similar to the role of w with time t; i.e., while w indicates the time harmonic nature of a function, k is its space harmonic nature. After observing (1)–(3), we may be tempted to construe that a wave is nothing but a DSS. Although every wave is a DSS, every DSS may not be a wave; Figure 3 and subsequent equations explain this. When we observe the difference between (4) and (5), both are DSSs but (5) is a wave and (4) is not.

$$V(x,y,t) = V_o(x,y)\cos(wt) \quad (4)$$

$$V(x,y,z,t) = V_o(x,y)\cos(wt\text{-}kz). \quad (5)$$

We may observe that none of the space variables in (4) appear in the angle of the sinusoid, whereas in (5) one of those does. The difference may be seen from another perspective, as shown in Figure 3, where a signal can become a wave only when at least one of its space axes is colocated with that of time.

What we have learned

So far, we clearly know the difference between the distributed amplitude of an entity and a signal. Keeping in view the lumped/point and distributed nature of amplitudes, PSSs and DSSs are defined. An important difference between waves and other signals is learned. Furthermore, a signal may or may not contain any information. Our preceding discussions and arguments usher us to reframe the definition of a signal with a greater technical precision. Consequently, a signal, represented as a function of one or more variables, may be defined as an observable change in a quantifiable entity.

Author

Pragnan Chakravorty (pciitkgp@ieee.org) received his B.E. degree in electronics engineering from Nagpur University, India, in 2001; his M.Tech. degree in radio-frequency and microwave engineering from the Indian Institute of Technology, Kharagpur, in 2005; and his Ph.D. degree in electronics and communication engineering from the National Institute of Technology, Durgapur, India, in 2017. Currently, he is a professor in the Department of Electronics and Telecommunication Engineering at Ramrao Adik Institute of Technology, Navi Mumbai, India. He is the founding director of Clique for Applied Research in Electronic Technology, a research trust.

References

[1] E. Gregersen. (2011, Apr.). Linear motion. [Online]. Available: https://www.britannica.com/science/linear-motion

[2] A. V. Oppenheim and A. S. Willsky, *Signals & Systems*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1997.

[3] B. P. Lathi, *Signal Processing and Linear Systems*. Berkeley, CA: Cambridge Press. 1998.

[4] S. S. Haykin and B. V. Veen, *Signals & Systems*. Hoboken, NJ: Wiley, 2003.

[5] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, July 1948.

[6] B. P. Lathi, *Modern Digital & Analog Communications Systems*, 2nd ed. London: Oxford Univ. Press, 1995

[7] S. Das, G. Ram, P. Chakravorty, D. Mandal, R. Kar, and S. P. Ghoshal, "Optimization of antenna arrays for SLL reduction towards Pareto objectivity using GA variants," in *Proc. IEEE Symp. Series Computational Intelligence*, Cape Town, 2015, pp. 1164–1169.

[8] R. F. Harrington, *Time-Harmonic Electromagnetic Fields*, Piscataway, NJ: IEEE Press, 2001.

[9] P. Chakravorty, "Analysis of rectangular waveguides: An intuitive approach," *IETE J. Educ.*, vol. 55, no. 2, pp. 76–80, 2014.

[10] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, Oct. 1948.

Yuejie Chi

Low-Rank Matrix Completion

Imagine one observes a small subset of entries in a large matrix and aims to recover the entire matrix. Without a priori knowledge of the matrix, this problem is highly ill-posed. Fortunately, data matrices often exhibit low-dimensional structures that can be used effectively to regularize the solution space. The celebrated effectiveness of principal component analysis (PCA) in science and engineering suggests that most variability of real-world data can be accounted for by projecting the data onto a few directions known as the principal components. Correspondingly, the data matrix can be modeled as a low-rank matrix, at least approximately. Is it possible to complete a partially observed matrix if its rank, i.e., its maximum number of linearly independent row or column vectors, is small?

Low-rank matrix completion arises in a variety of applications in recommendation systems, computer vision, and signal processing. As a motivating example, consider users' ratings of products arranged in a rating matrix. Each rating may only be affected by a small number of factors—such as price, quality, and utility—and how they are reflected on the products' specifications and users' expectations. Naturally, this suggests that the rating matrix is low rank, since the numbers of users and products are much higher than the number of factors. Often, the rating matrix is sparsely observed, and it is of great interest to predict the missing ratings to make targeted recommendations.

Relevance

The theory and algorithms of low-rank matrix completion have been significantly expanded in the last decade with converging efforts from signal processing, applied mathematics, statistics, optimization, and machine learning. This lecture note provides an introductory exposition of some key results in this rapidly developing field.

Prerequisites

We expect the readers to be familiar with basic concepts in linear algebra, optimization, and probability.

Problem statement

Let $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ be a rank-r matrix, whose thin singular value decomposition (SVD) is given as

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}, \tag{1}$$

where $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{V} \in \mathbb{R}^{n_2 \times r}$ are composed of orthonormal columns, and $\boldsymbol{\Sigma}$ is an r-dimensional diagonal matrix with the singular values arranged in a nonincreasing order, i.e., $\sigma_1 \geq \cdots \geq \sigma_r > 0$. The "degrees of freedom" of $\boldsymbol{M}$ is $(n_1 + n_2 - r)r$, which is the total number of parameters we need to uniquely specify $\boldsymbol{M}$.

Assume we are given partial observations of $\boldsymbol{M}$ over an index set $\Omega \subset \{1, 2, \ldots, n_1\} \times \{1, 2, \ldots, n_2\}$. To concisely put it, define the observation operator $\mathcal{P}_\Omega: \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ as

$$[\mathcal{P}_\Omega(\boldsymbol{M})]_{ij} = \begin{cases} M_{ij}, & (i,j) \in \Omega \\ 0, & \text{otherwise} \end{cases}.$$

Our goal is to recover $\boldsymbol{M}$ from $\mathcal{P}_\Omega(\boldsymbol{M})$, when the number of observation $m = |\Omega| \ll n_1 n_2$ is much smaller than the number of entries in $\boldsymbol{M}$, under the assumption that $\boldsymbol{M}$ is low rank, i.e., $r \ll \min\{n_1, n_2\}$. For notational simplicity in the sequel, let $n = \max\{n_1, n_2\}$.

Solution

Which low-rank matrices can we complete?

To begin, we ask the following question: What kind of low-rank matrices can we complete? As motivation, consider the following 4×4 rank-1 matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$, given as

$$\boldsymbol{M}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{M}_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

The matrix $\boldsymbol{M}_1$ is more difficult to complete, since most of its entries are zero, and we need to collect more measurements to make sure enough mass comes from its nonzero entries. In contrast, the mass of $\boldsymbol{M}_2$ is more uniformly distributed across all entries, making it easier to propagate information from one entry to another.

To put it differently, a low-rank matrix is easier to complete if its energy spreads evenly across different coordinates.

To put it differently, a low-rank matrix is easier to complete if its energy spreads evenly across different coordinates. This property is captured by the notion of *coherence* [1], which measures the alignment between the column/row spaces of the low-rank matrix with standard basis vectors. For a matrix $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r}$ with orthonormal columns, let $\boldsymbol{P}_U$ be the orthogonal projection onto the column space of $\boldsymbol{U}$. The coherence parameter of $\boldsymbol{U}$ is defined as

Digital Object Identifier 10.1109/MSP.2018.2832197
Date of publication: 28 August 2018

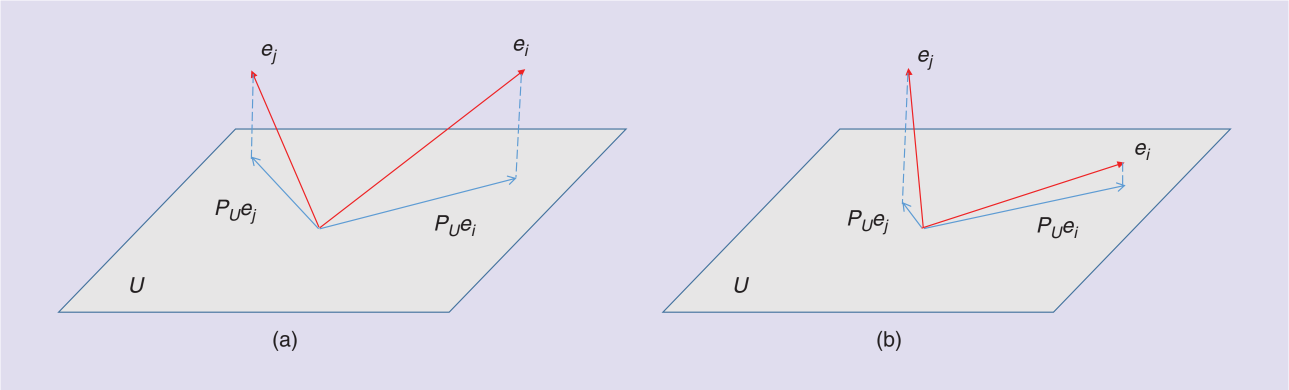


FIGURE 1. An illustration of the coherence parameter $\mu(\boldsymbol{U})$. $\mu(\boldsymbol{U})$ is small when all the standard basis vectors $\boldsymbol{e}_i$ have approximately the same projections onto the subspace $\boldsymbol{U}$, as shown in (a); $\mu(\boldsymbol{U})$ is large if $\boldsymbol{U}$ is too aligned with certain standard basis vector, as shown in (b).

$$\begin{aligned}\mu(\boldsymbol{U}) &= \frac{n_1}{r}\max_{1\le i\le n_1}\|\boldsymbol{P}_U\boldsymbol{e}_i\|_2^2\\ &= \frac{n_1}{r}\max_{1\le i\le n_1}\|\boldsymbol{U}^\top\boldsymbol{e}_i\|_2^2, \qquad (2)\end{aligned}$$

where $\boldsymbol{e}_i$ is the ith standard basis vector. Figure 1 provides a geometric illustration of the coherence parameter $\mu(\boldsymbol{U})$.

For a low-matrix $\boldsymbol{M}$ whose SVD is given in (1), the coherence of $\boldsymbol{M}$ is defined as

$$\mu = \max\{\mu(\boldsymbol{U}), \mu(\boldsymbol{V})\}. \qquad (3)$$

Notably, the coherence μ is determined by the singular vectors of $\boldsymbol{M}$ and independent of its singular values. Since $1 \le \mu(\boldsymbol{U}) \le n_1/r$ and $1 \le \mu(\boldsymbol{V}) \le n_2/r$, we have $1 \le \mu \le n/r$. In the previous example, the coherence of $\boldsymbol{M}_1$ matches the upper bound n/r, while the coherence of $\boldsymbol{M}_2$ matches the lower bound one. The smaller μ is, the easier it is to complete the matrix.

Which observation patterns can we handle?

Low-rank matrix completion can still be hopeless even when most of the entries are revealed. Consider, for example, the following observation pattern for a 4×4 matrix:

$$\begin{bmatrix} \star & \star & \star & ? \\ \star & \star & \star & ? \\ \star & \star & \star & ? \\ \star & \star & \star & ? \end{bmatrix},$$

where $\star$ indicates an observed entry, and ? indicates a missing entry. The last column of the matrix cannot be recovered since it can lie anywhere in the column space of the low-rank matrix. Therefore, we require at least r observations per column/row. To bypass such pessimistic observation patterns, it is useful to think of random observation patterns. A popular choice is the Bernoulli model, where each entry is observed independently and identically with probability $p := m/(n_1 n_2)$. By a coupon-collecting argument [2], under the Bernoulli model, it is impossible to recover a low-rank matrix with less than some constant times $\mu nr \log n$ measurements using any algorithm, which is referred to as the *information-theoretic lower bound*. Compared with the degrees of freedom, which is on the order of nr, we pay a price in sample complexity by a factor of $\mu \log n$, highlighting again the role of coherence in low-rank matrix completion.

Matrix completion via convex optimization

We present the first algorithm based on convex optimization. To promote the low-rank structure of the solution, a natural heuristic is to find the matrix with the minimum rank that is consistent with the observations, leading to

$$\begin{aligned}&\min_{\boldsymbol{\Phi}\in\mathbb{R}^{n_1\times n_2}} \operatorname{rank}(\boldsymbol{\Phi})\\ &\text{s.t.}\quad \mathcal{P}_\Omega(\boldsymbol{\Phi}) = \mathcal{P}_\Omega(\boldsymbol{M}). \qquad (4)\end{aligned}$$

However, since rank minimization is NP-hard, the above formulation is intractable. Motivated by the success of ℓ_1 norm minimization for sparse recovery in compressed sensing [3], we consider convex relaxation for the rank heuristic. Observing that the rank of $\boldsymbol{\Phi}$ equals to the number of its nonzero singular values, we replace rank ($\boldsymbol{\Phi}$) by the sum of its singular values, denoted as the nuclear norm:

$$\|\boldsymbol{\Phi}\|_* \triangleq \sum_{i=1}^{\min\{n_1,n_2\}} \sigma_i(\boldsymbol{\Phi}),$$

where $\sigma_i(\boldsymbol{\Phi})$ is the ith singular value of $\boldsymbol{\Phi}$. The nuclear norm is the tightest convex relaxation of the rank constraint, i.e., the nuclear norm ball $\{\boldsymbol{\Phi} : \|\boldsymbol{\Phi}\|_* \le 1\}$ is the convex hull of the collection of unit-norm rank-1 matrices: $\{\boldsymbol{u}\boldsymbol{v}^\top : \|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1\}$. Notably, the nuclear norm is also unitarily invariant, and can be represented as the solution to a semidefinite program,

$$\begin{aligned}\|\boldsymbol{\Phi}\|_* &= \min_{\boldsymbol{W}_1,\boldsymbol{W}_2} \frac{1}{2}(\operatorname{Tr}(\boldsymbol{W}_1) + \operatorname{Tr}(\boldsymbol{W}_2))\\ &\text{s.t.}\quad \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{\Phi} \\ \boldsymbol{\Phi}^\top & \boldsymbol{W}_2 \end{bmatrix} \succeq 0.\end{aligned}$$

Hence, instead of solving (4) directly, we solve nuclear norm minimization, which searches for a matrix with the minimum nuclear norm that satisfies all the measurements:

$$\min_{\boldsymbol{\Phi}\in\mathbb{R}^{n_1\times n_2}} \|\boldsymbol{\Phi}\|_* \quad \text{s.t.}\quad \mathcal{P}_\Omega(\boldsymbol{\Phi}) = \mathcal{P}_\Omega(\boldsymbol{M}). \qquad (5)$$

This gives a convex program that can be solved efficiently in polynomial time. Moreover, it doesn't require knowledge of the rank a priori.

The performance of nuclear norm minimization has been investigated in a recent line of elegant works [2]–[5], which suggests it can exactly recover a low-rank matrix as soon as the number

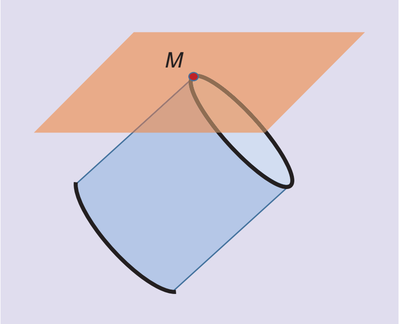


FIGURE 2. A geometric illustration of nuclear norm minimization: the cylinder represents level sets of the nuclear norm, and the hyperplane represents the measurement constraint. The two sets intersect at the thickened edges, which correspond to low-rank solutions.

of measurements is slightly larger than the information-theoretic lower bound by a logarithmic factor. Suppose that each entry of $\boldsymbol{M}$ is observed independently with probability $p \in (0,1)$. If p satisfies

$$p \geq C \frac{\mu r \log^2 n}{n},$$

for some large enough constant $C > 0$, then with high probability, the nuclear norm minimization algorithm (5) exactly recovers $\boldsymbol{M}$ as the unique optimal solution of (5). Figure 2 illustrates the geometry of nuclear norm minimization when the number of measurements is sufficiently large. When both μ and r are much smaller than n, this means we can recover a low-rank matrix even when the proportion of observations is vanishingly small.

Matrix completion via nonconvex optimization

The computational and memory complexities of nuclear norm minimization can be quite expensive for large-scale problems, even with first-order methods, due to optimizing over and storing the matrix variable $\boldsymbol{\Phi}$. Therefore, it is necessary to consider alternative approaches whose complexities scale more favorably in n. This leads to the second algorithm based on gradient descent using a proper initialization. If the rank of the matrix $\boldsymbol{M}$ is known, it is natural to incorporate this knowledge and consider a rank-constrained least-squares problem

$$\min_{\boldsymbol{\Phi} \in \mathbb{R}^{n_1 \times n_2}} \| \mathcal{P}_\Omega(\boldsymbol{\Phi} - \boldsymbol{M}) \|_{\mathrm{F}}^2, \quad \text{s.t rank } (\boldsymbol{\Phi}) \leq r, \tag{6}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm of a matrix. Invoking the low-rank factorization $\boldsymbol{\Phi} = \boldsymbol{X}\boldsymbol{Y}^\top$, where $\boldsymbol{X} \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{Y} \in \mathbb{R}^{n_2 \times r}$, we can rewrite (6) as an unconstrained, yet nonconvex optimization problem:

$$\min_{\boldsymbol{X},\boldsymbol{Y}} f(\boldsymbol{X},\boldsymbol{Y}) := \| \mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}) \|_{\mathrm{F}}^2. \tag{7}$$

On one end, the memory complexities of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are linear in n instead of quadratic in n when dealing with $\boldsymbol{\Phi}$. On the other end, we can only determine $\boldsymbol{X}$ and $\boldsymbol{Y}$ up to invertible transforms in (7), since for any invertible matrix $\boldsymbol{Q} \in \mathbb{R}^{r \times r}$, we have $\boldsymbol{X}\boldsymbol{Y}^\top = (\boldsymbol{X}\boldsymbol{Q})(\boldsymbol{Y}\boldsymbol{Q}^{-\top})^\top$. To fix the scaling ambiguity, it is useful to consider a modified loss function

$$F(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{4p} f(\boldsymbol{X},\boldsymbol{Y}) + \frac{1}{16} \| X^\top \boldsymbol{X} - \boldsymbol{Y}^\top Y \|_{\mathrm{F}}^2,$$

where the second term is introduced to motivate solutions where $\boldsymbol{X}$ and $\boldsymbol{Y}$ have balanced norms. The observation probability p, if not known, can be faithfully estimated by the sample proportion $|\Omega|/(n_1 n_2)$.

How do we optimize the nonconvex loss $F(\boldsymbol{X},\boldsymbol{Y})$? A plausible strategy proceeds in two steps.

1) The first step aims to find an initialization that is close to the ground truth, which can be provided via the so-called spectral method [6]. Consider the partially observed matrix $(1/p)\mathcal{P}_\Omega(\boldsymbol{M})$, which is an unbiased estimate of $\boldsymbol{M}$ with expectation $\mathbb{E}[(1/p)\mathcal{P}_\Omega(\boldsymbol{M})] = \boldsymbol{M}$. Therefore, its best rank-r approximation produces a reasonably good initial guess. Let $\boldsymbol{U}_0 \boldsymbol{\Sigma}_0 \boldsymbol{V}_0^\top$ be the best rank-r approximation of $(1/p)\mathcal{P}_\Omega(\boldsymbol{M})$, where $\boldsymbol{U}_0 \in \mathbb{R}^{n_1 \times r}, \boldsymbol{V}_0 \in \mathbb{R}^{n_2 \times r}$ contain orthonormal columns and $\boldsymbol{\Sigma}_0$ is an $r \times r$ diagonal matrix. The spectral initialization sets $\boldsymbol{X}_0 = \boldsymbol{U}_0 \boldsymbol{\Sigma}_0^{1/2}$ and $\boldsymbol{Y}_0 = \boldsymbol{V}_0 \boldsymbol{\Sigma}_0^{1/2}$.
2) The second step aims to refine the initial estimate locally via simple iterative methods, such as gradient descent [7], [8], following the update rule

$$\begin{bmatrix} \boldsymbol{X}_{t+1} \\ \boldsymbol{Y}_{t+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t \end{bmatrix} - \eta_t \begin{bmatrix} \nabla_{\boldsymbol{X}} F(\boldsymbol{X}_t, \boldsymbol{Y}_t) \\ \nabla_{\boldsymbol{Y}} F(\boldsymbol{X}_t, \boldsymbol{Y}_t) \end{bmatrix}, \tag{8}$$

where η_t is the step size, and $\nabla_{\boldsymbol{X}} F(\boldsymbol{X},\boldsymbol{Y}), \nabla_{\boldsymbol{Y}} F(\boldsymbol{X},\boldsymbol{Y})$ are the partial derivatives with respect to $\boldsymbol{X}$ and $\boldsymbol{Y}$ that can be derived easily.

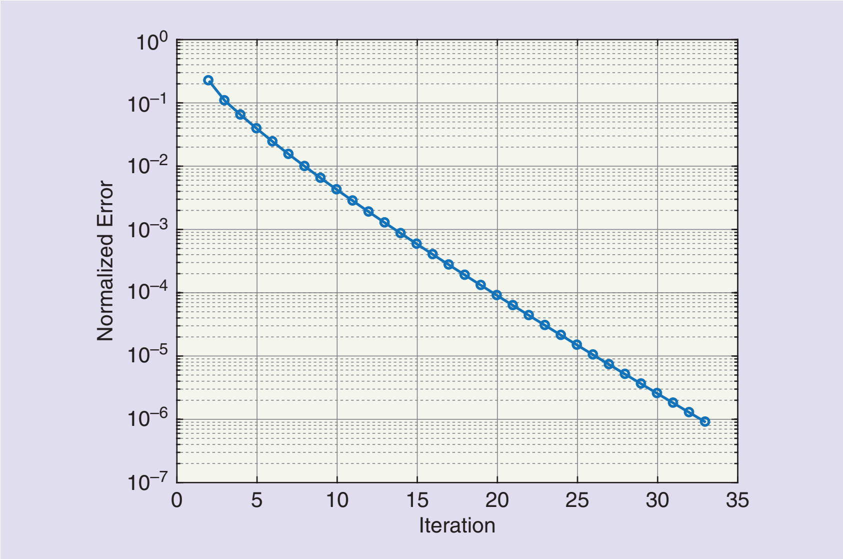


FIGURE 3. The normalized error of low-rank matrix completion with respect to the iteration count via gradient descent with the spectral initialization for a $10^4 \times 10^4$ matrix of rank-10 using about 5% observations.

Recall the SVD of $\boldsymbol{M}$ in (1), and denote $\boldsymbol{X}^\natural = \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{Y}^\natural = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}$; this allows us to write the factorization as $\boldsymbol{M} = \boldsymbol{X}^\natural \boldsymbol{Y}^{\natural\top}$ and call $\boldsymbol{Z}^\natural = [\boldsymbol{X}^{\natural\top}, \boldsymbol{Y}^{\natural\top}]^\top \in \mathbb{R}^{(n_1+n_2)\times r}$ the ground truth. Since $\boldsymbol{Z}^\natural$ is only identifiable up to orthonormal transforms, let the optimal transform between the tth iterate $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top \in \mathbb{R}^{(n_1+n_2)\times r}$ and $\boldsymbol{Z}^\natural$ as

$$\boldsymbol{H}_t := \underset{\boldsymbol{R}\in\mathbb{R}^{r\times r}, \boldsymbol{R}\boldsymbol{R}^\top = \boldsymbol{I}}{\operatorname{argmin}} \left\| \boldsymbol{Z}_t \boldsymbol{R} - \boldsymbol{Z}^\natural \right\|_{\mathrm{F}}.$$

Assume the condition number $\kappa := \sigma_1/\sigma_r$ of $\boldsymbol{M}$ is a bounded constant, then as long as

$$p \geq C_1 \frac{\mu^3 r^3 \log^3 n}{n}$$

for some sufficiently large constant $C_1 > 0$, with high probability, the iterates satisfy [8]

$$\| \boldsymbol{Z}_t \boldsymbol{H}_t - \boldsymbol{Z}^\natural \|_{\mathrm{F}} \leq C_2 \rho^t \mu r \frac{1}{\sqrt{np}} \| \boldsymbol{Z}^\natural \|_{\mathrm{F}}, \quad \forall t \geq 0,$$

where $C_2 > 0, 0 < \rho < 1$ are some constants, provided that the step size $0 < \eta_t \equiv \eta \leq 2/(25\kappa\sigma_1)$. Hence, gradient descent converges at a geometric rate, as soon as the number of measurements is on the order of $\mu^3 r^3 n \log^3 n$, which scales linearly in n up to logarithmic factors. To reach ϵ-accuracy, i.e., $\|\boldsymbol{Z}_t\boldsymbol{H}_t - \boldsymbol{Z}^\natural\|_{\mathrm{F}} / \|\boldsymbol{Z}^\natural\|_{\mathrm{F}} \leq \epsilon$, gradient descent needs an order of $\log(1/\epsilon)$ iterations. The number of iterations is independent of the problem size and therefore the computational cost is much cheaper in conjunction with low cost per iteration.

Summary

Table 1 summarizes the figures-of-merit of the discussed algorithms using state-of-the-art theory.

Numerical example

Let $\boldsymbol{M}$ be a rank-10 matrix of size $10^4 \times 10^4$ with about 5% of observed entries, i.e., $p = 0.05$, where $\boldsymbol{X}^\natural$ and $\boldsymbol{Y}^\natural$ are generated with i.i.d. standard Gaussian entries. We implement gradient descent with spectral initialization to recover $\boldsymbol{M}$. Figure 3 plots the normalized error $\|\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}\|_{\mathrm{F}} / \|\boldsymbol{M}\|_{\mathrm{F}}$ with respect to the iteration counts, which verifies the geometric convergence predicted by the theory. Indeed, the normalized error is below 10^{-5} within 30 iterations!

Table 1. Figure-of-merits for low-rank matrix completion in terms of order-wise sample complexity and computational complexity.

	Sample Complexity	Computational Complexity
Information-theoretic lower bound	$\mu n r \log n$	NP-hard
Nuclear norm minimization	$\mu n r \log^2 n$	Polynomial time
Gradient descent with spectral initialization	$\mu^3 n r^3 \log^3 n$	Linear time

What we have learned

Under mild statistical models, low-rank matrix completion admits efficient algorithms with provable near-optimal performance guarantees, using both convex and nonconvex optimization techniques. The theory and algorithms discussed herein can be extended to recover matrices that are approximately low rank using noisy measurements. Low-rank matrix completion can be viewed as a special case of low-rank matrix estimation using an underdetermined set of linear equations. Other linear measurement patterns are also actively studied, motivated by applications such as sensor network localization, phase retrieval, quantum state tomography, and so on. Furthermore, low-rank matrix completion can be made robust even when many of the observations are corrupted by outliers of arbitrary magnitudes, known as the *sparse* and *low-rank decomposition problem* [9].

Low-rank structures are ubiquitous in modern data science problems and becoming increasingly popular as a modeling tool. Understanding the algorithmic and theoretical properties of estimation of low-rank structures is still an active area of research that will have a growing impact in future years. For a recent survey on low-rank matrix estimation, please see [10].

Acknowledgments

This work was supported by Air Force Office of Scientific Research under FA9550-15-1-0205, the Office of Naval Research under N00014-18-1-2142, and the National Science Foundation under CAREER ECCS-1818571 and CCF-1806154.

Author

Yuejie Chi (yuejiechi@cmu.edu) received her B.E. (Hon.) degree in electrical engineering from Tsinghua University, Beijing, China, in 2007 and her Ph.D. degree in electrical engineering from Princeton University, New Jersey, in 2012. She is currently an associate professor with the Department of Electrical and Computer Engineering at Carnegie Mellon University, Pittsburgh, Pennsylvania. Her research interests include statistical signal processing, machine learning, and large-scale optimization and their applications in data science, inverse problems, imaging, and sensing systems. She is a Senior Member of the IEEE.

References

[1] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, Apr. 2009.

[2] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[3] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[4] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1548–1566, Mar. 2011.

[5] Y. Chen, "Incoherence-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2909–2923, 2015.

[6] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[7] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6535–6579, 2016.

[8] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution," arXiv Preprint, arXiv:1711.10467, 2017.

[9] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 11, 2011.

[10] Y. Chen and Y. Chi, "Harnessing structures in big data via guaranteed low-rank matrix estimation," *IEEE Signal Process. Mag.*, vol. 35, no. 4, pp. 14–31, July 2018.

SP

DATES AHEAD

Please send calendar submissions to:
Dates Ahead, Att: Jessica Welsh, E-mail: j.welsh@ieee.org

2018

SEPTEMBER

26th European Signal Processing Conference (EUSIPCO)
3–7 September, Rome, Italy.
General Chair: Patrizio Campisi
General Cochair: Josef Kittler
URL: http://www.eusipco2018.org/

IEEE International Conference on Content-Based Multimedia Indexing (CBMI)
4–6 September, La Rochelle, France.
General Chair: Renaud Péteri
URL: http://cbmi2018.univ-lr.fr/

IEEE International Workshop on Machine Learning for Signal Processing (MLSP)
17–20 September, Aalborg, Denmark.
General Chair: Zheng-Hua Tan
URL: http://mlsp2018.conwiz.dk/home.htm

16th International Workshop on Acoustic Signal Enhancement (IWAENC)
17–20 September, Tokyo, Japan.
General Chairs: Hiroshi Saruwatari and Shoji Makino
URL: http://www.iwaenc2018.org/

OCTOBER

25th IEEE International Conference on Image Processing (ICIP)
7–10 October, Athens, Greece.
General Chairs: Christophoros Nikou and Kostas Plataniotis
URL: https://2018.ieeeicip.org

IEEE Workshop on Signal Processing Systems (SiPS)
21–24 October, Cape Town, South Africa.
General Chair: Tokunbo Ogunfunmi
URL: http://www.sips2018.org/

Asilomar Conference on Signals, Systems, and Computers (ACSSC)
28–31 October, Pacific Grove, California, United States.
General Chair: Visa Koivunen
URL: http://www.asilomarsscconf.org/

Digital Object Identifier 10.1109/MSP.2018.2860218
Date of publication: 28 August 2018

©ISTOCKPHOTO.COM/GIVAGA

The IEEE International Symposium on Biomedical Imaging will be held 8–11 April 2019 in beautiful Venice, Italy.

NOVEMBER

Tenth Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA 2018)
12–15 November, Honolulu, Hawaii, United States.
General Chairs: Yih-Fang Huang, Anthony Kuh, and Susanto Rahardja
URL: https://apsipa2018.org

Sixth IEEE Global Conference on Signal and Information Processing (GlobalSIP)
26–28 November, Anaheim, California, United States.
General Chairs: Shuguang Cui and Hamid Jafarkhani
URL: http://2018.ieeeglobalsip.org/

15th IEEE International Conference on Advanced Video and Signals-Based Surveillance (AVSS)
27–30 November, Auckland, New Zealand.
General Chairs: Reinhard Klette and Mohan Kankanhalli
URL: https://avss2018.org

DECEMBER

2018 IEEE International Workshop on Information Forensics and Security (WIFS)
11–13 December, Hong Kong.
General Chair: Ajay Kumar
URL: https://wifs2018.comp.polyu.edu.hk/

2018 IEEE Spoken Language Technology Workshop (SLT)
18–21 December, Athens, Greece.
Cochairs: Vangelis Karkaletsis, Yannis Stylianou, and Srinivas Bangalore
URL: http://www.slt2018.org

2019

APRIL

IEEE International Symposium on Biomedical Imaging (ISBI)
8–11 April, Venice, Italy.
General Chairs: Marius George Linguraru and Enrico Grisan
URL: https://biomedicalimaging.org/2019/

MAY

44th IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)
12–17 May, Brighton, United Kingdom.
General Chairs: Saeid Sanei and Lajos Hanzo
URL: http://icassp2019.com

JULY

IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)
2–5 July, Cannes, France.
General Chair: David Gesbert
URL: http://www.spawc2019.org/ SP

(continued from page 188)

skepticism and is very much on the way, and fast.

The current rush for low-hanging fruits that are ripe for ML will eventually slow. When it does, a general consensus of where—and for what purpose—the data-driven ML techniques are appropriate will emerge. Already, and even more so at that stage, it is and will be important to innovate in the space between the established communication system models with provably optimal solutions and purely data-driven methods. One advantage of the SPCOM area is that well-developed models exist based on the physics of electromagnetic propagation. Our research community has historically invested significant efforts to refine stochastic channel models and develop software packages for simulating the full communication chain, including the more complex ray-tracing simulators of the wireless environment. Yet, we are all keenly aware that all models have their limitations; thus an interesting future direction will be to make use of these models in conjunction with the data-driven approaches, not only for the testing and evaluation of proposed solutions but also for data generation.

The SPCOM-TC very much welcomes SPS members from diverse backgrounds to participate in our technical activities. In particular, research within many signal processing research communities has been accelerated by the creation of open and easily accessible software tools. The SPCOM community could be similarly helped by common and open simulators of complex communication systems, with a number of predefined scenarios that make it easy to get research started and allow for ready comparisons of competing solutions. While this need has existed previously, it will be exacerbated by the proliferation of data-driven approaches. It will also be necessary to raise the scientific quality of such works, in particular when it comes to reproducibility.

We invite SPS members to get involved by signing up as affiliated members of SPCOM-TC from the SPS website (https://signalprocessingsociety.org/get-involved/signal-processing-communications-and-networking). The TC membership election takes place in October every year. We sincerely hope to continue this discussion on the future technical directions of our TC with many of you in an intellectually stimulating environment at our annual workshop SPAWC or at the next IEEE International Conference on Acoustics, Speech, and Signal Processing.

Authors

Wei Yu (weiyu@ece.utoronto.ca) received his B.A.Sc. degree in computer engineering and mathematics from the University of Waterloo, Canada, in 1997, and his Ph.D. degree in electrical engineering from Stanford University, California, in 2002. He is currently a professor and the Canada Research Chair in Information Theory and Wireless Communications at the University of Toronto, Canada. He received the IEEE Signal Processing Society Best Paper Award in 2008 and 2017. He is a Fellow of the IEEE and the Canadian Academy of Engineering. He currently serves as the chair of the IEEE Signal Processing for Communications and Networking Technical Committee.

Joakim Jaldén (jalden@kth.se) received his M.Sc. and Ph.D. degrees in electrical engineering from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2002 and 2007, respectively. After a postdoctoral position at the Vienna University of Technology, Austria, he returned to KTH where he is now a professor of signal processing. He has been a member of the IEEE Signal Processing for Communications and Networking Technical Committee since 2013 and currently serves as its vice-chair.

Martin Haardt, Christoph Mecklenbräuker, and Peter Willett

Highlights from the Sensor Array and Multichannel Technical Committee

Spotlight on the IEEE Signal Processing Society Technical Committees

The IEEE Signal Processing Society Sensor Array and Multichannel Technical Committee (SAM TC) promotes activities within the technical areas of sensor array processing and multichannel statistical signal processing, including

- beamforming
- direction-of-arrival estimation
- source localization
- multiple-input multiple-output (MIMO) systems
- compressed sensing
- sparse modeling
- tensor-based signal processing
- deep neural networks
- machine learning for sensor arrays
- signal processing for sensor networks
- network beamforming
- blind source separation
- channel identification
- array processing for radar, sonar, communications, microphone arrays, and biomedical applications.

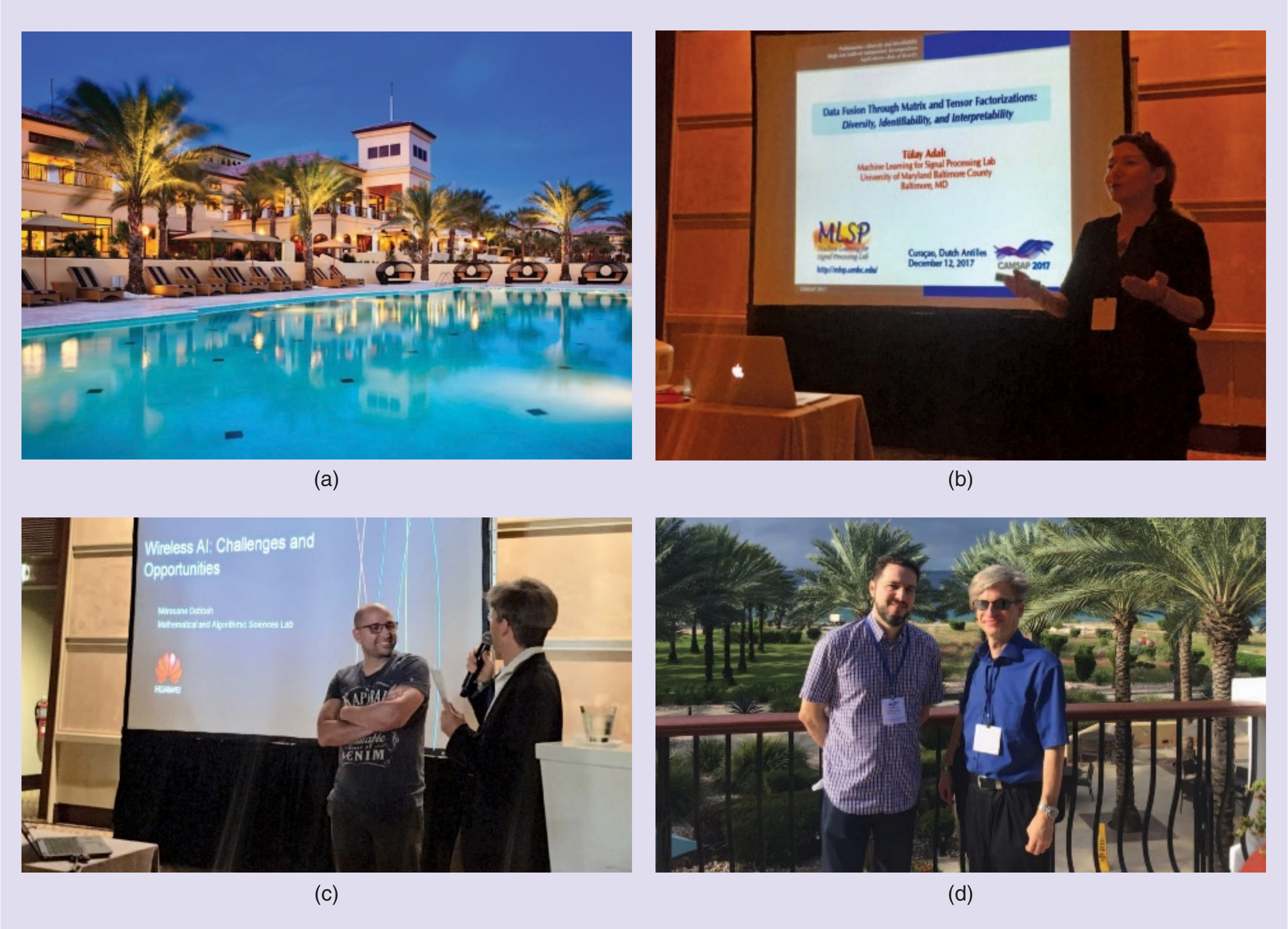
Digital Object Identifier 10.1109/MSP.2018.2835718
Date of publication: 28 August 2018

Table 1. Statistics of recent CAMSAP workshops (regular and special session papers).

Year	Location	Number of Accepted Papers	Attendance
2009	Aruba, Dutch Antilles	103	114
2011	San Juan, Puerto Rico	133	124
2013	Saint Martin, French Antilles	125	140
2015	Cancun, Mexico	136	142
2017	Curaçao, Dutch Antilles	168	196

Our biannual IEEE International Workshop on Computational Advances in Multisensor Adaptive Processing (CAMSAP) has been organized in December every odd-numbered year since 2005. The statistics of recent editions are summarized in Table 1. The seventh edition was held 10–13 December 2017 at the Santa Barbara Beach and Golf Resort, Curaçao, Dutch Antilles (https://signalprocessingsociety.org/CAMSAP2017/). After Josef A. Nossek and Georgios B. Giannakis delivered instructive tutorials on 10 December, we had six exciting plenary presentations (given by Yonina C. Eldar, Daniel P. Palomar, Antonio Ortega, Tülay Adalı, Nikos Sidiropoulos, and our plenary speaker from industry Mérouane Debbah), 19 special invited sessions, and ten regular sessions. As indicated in Table 1, CAMSAP 2017 received the highest number of submissions thus far and attracted a record number of attendees—38% of whom were student participants. Some photos from the workshop are shown in Figure 1. An IEEE SPS questionnaire was distributed to all participants to evaluate how satisfied they were with the workshop. The results were extremely positive. CAMSAP 2019 will be held in December in Guadeloupe, French West Indies.

The Tenth IEEE Sensor Array and Multichannel Signal Processing (SAM) Workshop took place 8–11 July 2018 in Sheffield, United Kingdom (http://www.sam2018.group.shef.ac.uk/). SAM is a biannual series of workshops that takes place midyear of even-numbered years. The ninth edition occurred 10–13 July 2016 in Rio de Janeiro, Brazil, about a month before the 2016 Summer Olympic Games. This ideal timing ensured


Data Fusion Through Matrix and Tensor Factorizations:
Diversity, Identifiability, and Interpretability
Tülay Adalı
Machine Learning for Signal Processing Lab
University of Maryland Baltimore County
Baltimore, MD
MLSP
http://mlsp.umbc.edu/
Curaçao, Dutch Antilles
December 12, 2017
CAMSAP 2017
Wireless AI: Challenges and Opportunities
Mérouane Debbah
Mathematical and Algorithmic Sciences Lab
HUAWEI

(a)

(b)

(c)

(d)

FIGURE 1. Impressions from CAMSAP 2017 in Curaçao, Dutch Antilles. (a) The conference venue, the Santa Barbara Beach and Golf Resort Curaçao. (b) The plenary presentation by Tülay Adalı. (c) Mérouane Debbah answers questions after his plenary presentation. (d) (From left) General chairs of CAMSAP 2017 André de Almeida and Martin Haardt.

that plenty of cultural activities and festivities (in addition to the excellent scientific highlights) made SAM 2016 a memorable experience. The statistics of the last five editions of SAM are summarized in Table 2. SAM 2016 received a record number of submissions and the highest number of attendees. Two very attractive proposals (from China and Mexico) for the 11th SAM Workshop in 2020 were presented at the recent SAM TC meeting in Calgary, Canada. For the first time in its history, the SAM workshop will be organized in Asia, in Hangzhou, China, 8–11 June 2020 (http://www.isee.zju.edu.cn/sam2020/).

Table 2. Statistics of recent SAM workshops (regular and special session papers).

Year	Location	Number of Accepted Papers	Attendance
2008	Darmstadt, Germany	118	124
2010	Kibbutz Ma'ale Hahamisha, Israel	68	80
2012	Hoboken, New Jersey, United States	136	171
2014	A Coruña, Spain	134	161
2016	Rio de Janeiro, Brazil	156	215

Recent joint activities of the SAM TC and the SPS Audio and Acoustic Signal Processing TC include a joint special session, "Speaker Localization in Dynamic Real-Life Environments," at the 2017 International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and a special issue in *IEEE Journal of Selected Topics in Signal Processing* on acoustic source localization and tracking in dynamic real-life scenes.

In addition to our regular lecture and poster sessions, we also organized a panel discussion, "An Industry Perspective on Emerging Signal Processing Challenges," at ICASSP 2018 that will be summarized in a future issue of *IEEE Signal Processing Magazine*.

Hot topics within the technical fields of sensor array processing and multichannel statistical signal processing include sensor fusion, machine learning, tensor-based processing for multidimensional data, sensor data integrity, and security. New challenges are posed by upcoming applications in the power grid, mobility of people and goods, and massive multichannel signal processing as they appear in biomedical applications and fifth-generation wireless connectivity. This is a fast-moving arena with ample opportunities.

If you are interested in the activities of the SAM TC, please register as an affiliate member on our webpage (https://signalprocessingsociety.org/get-involved/sensor-array-and-multichannel).

Authors

Martin Haardt (Martin.Haardt@tu-ilmenau.de) is a professor and head of the Communications Research Laboratory at TU Ilmenau, Germany. He served as the Sensor Array and Multichannel Technical Committee chair in 2017 and 2018.

Christoph Mecklenbräuker (cfm@nt.tuwien.ac.at) is a professor of flexible wireless systems at TU Wien, Austria. He has served as the Sensor Array and Multichannel Technical Committee (SAM TC) vice chair since 2017 and will serve as the SAM TC chair in 2019 and 2020.

Peter Willett (peter.willett@uconn.edu) is a professor at the University of Connecticut, Storrs. He served as the Sensor Array and Multichannel Technical Committee chair in 2015 and 2016.

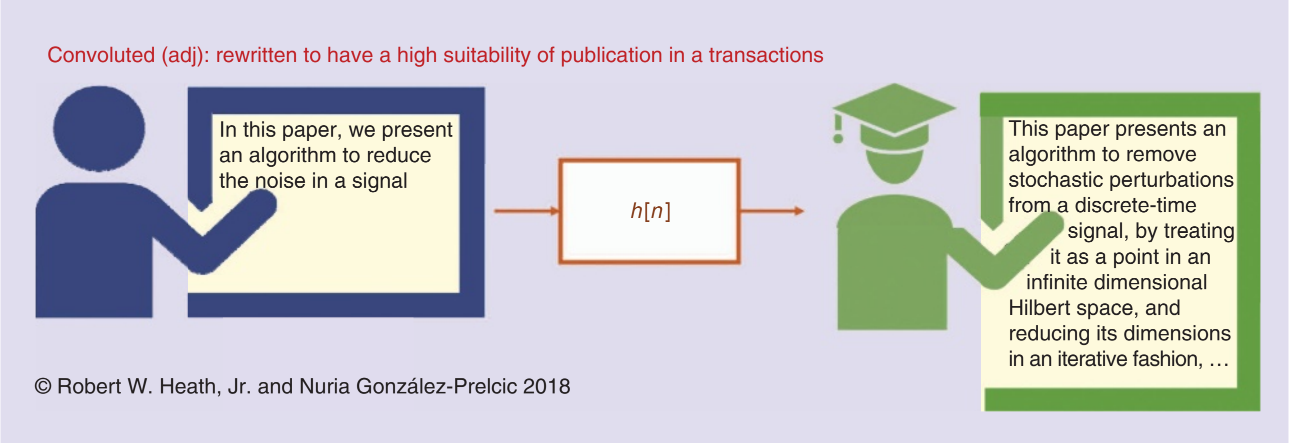
We want to hear from you!
Do you like what you're reading?
Your feedback is important.
Let us know—send the editor-in-chief an e-mail!
IEEE

IMAGE LICENSED BY GRAPHIC STOCK

HUMOR

Convoluted

by Robert W. Heath, Jr. and Nuria González-Prelcic



Convoluted (adj): rewritten to have a high suitability of publication in a transactions

In this paper, we present an algorithm to reduce the noise in a signal
$h[n]$
This paper presents an algorithm to remove stochastic perturbations from a discrete-time signal, by treating it as a point in an infinite dimensional Hilbert space, and reducing its dimensions in an iterative fashion, …

Digital Object Identifier 10.1109/MSP.2018.2858444
Date of publication: 28 August 2018

FOR YOUR CONSIDERATION

The authors of [1], which was published in the November 2017 issue of *IEEE Signal Processing Magazine*, wish to add an acknowledgment to their article. The acknowledgment is as follows:

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (number 2016R1A2B2014525) and by a grant from the National Science Foundation (IIS-1116656) awarded to Alan C. Bovik.

Digital Object Identifier 10.1109/MSP.2018.2856218
Date of publication: 28 August 2018

Reference

[1] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130–141, Nov. 2017.

ADVERTISING & SALES

© GRAPHIC STOCK

The Advertisers Index contained in this issue is compiled as a service to our readers and advertisers: the publisher is not liable for errors or omissions although every effort is made to ensure its accuracy. Be sure to let our advertisers know you found them through *IEEE Signal Processing Magazine.*

IEEE SIGNAL PROCESSING MAGAZINE REPRESENTATIVE
Mark David, Director, Business Development — Media & Advertising, Phone: +1 732 465 6473, Fax: +1 732 981 1855, m.david@ieee.org

COMPANY	PAGE NUMBER	WEBSITE	PHONE
MathWorks	CVR 4	www.mathworks.com/wireless	

Digital Object Identifier 10.1109/MSP.2017.2770716

Wei Yu and Joakim Jaldén

Perspectives in Signal Processing for Communications and Networking

Spotlight on the IEEE Signal Processing Society Technical Committees

The Signal Processing for Communications and Networking Technical Committee (SPCOM-TC) is one of the 12 technical committees (TCs) in the IEEE Signal Processing Society (SPS). Our mandate covers all technical areas in communication engineering and network science, including

- information transmission and reception
- channel modeling and estimation
- source and channel coding
- multicarrier and multiple-access communications
- array signal processing
- synchronization
- localization

as well as security; privacy; signal processing aspects of sensor and ad-hoc networks; cognitive radio systems; and distributed sensing, detection, estimation, and inference problems over the networks. The application areas range from terrestrial wireless systems to wireline, underwater, satellite, backscattering, and visible light communications, as well as on futuristic areas such as molecular, chemical, biological, and quantum communications. Our technical interests are tightly intertwined with that of the IEEE Communications Society and the IEEE Information Theory Society. We are committed to exploring the connections and cross-fertilization between these rapidly growing fields.

The SPCOM-TC organizes the IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC) each year in a unique all-poster format together with plenary and invited talks. The workshop has attracted increasing interest from the research community in recent years. The record of 268 attendees set at the last year's workshop in Sapporo, Japan, was almost matched by this year's SPAWC in Kalamata, Greece. In 2019, SPAWC will be held in Nice, France, and in 2020 in Atlanta, Georgia, United States.

Many of the classical topics of the TC, such as optimization-based solutions for the physical layer of wireless networks, are still gathering interest and producing novel theoretical results. However, it is their applications to newer settings such as millimeter-wave (mmWave) communication in the high-frequency band; massive multiple-input, multiple-output (MIMO) systems; and cooperative cloud radio-access networks that are now inspiring the most significant innovations. Much of these recent research activities have been driven by the emerging fifth generation (5G) wireless cellular standardization process with enhanced mobile broadband at target peak rates beyond tens of gigabits per second as its first stated goal. Efforts in utilizing the significantly larger bandwidth in the mmWave frequency band, in taking advantage of the potentially hundreds of spatial dimensions brought by the massive MIMO systems while accounting for their hardware limitations, and in cooperative signal processing to mitigate and cancel the dominant intercell interference, are the keys for the successful realization of 5G.

The 5G evolution is much more than just enhancing the data rate. The future Internet of Things calls for new use cases involving machine-type communications, particularly for meeting the challenges of connecting the large number of sensors and actuators, and for providing ultra-reliability and low-latency communications. These new requirements are driven by myriad vertical markets for the wireless technology, from industrial automation to remote health care, robotics, and autonomous driving, extending further to, e.g., communications and control of unmanned aerial vehicles and high-speed Internet service provision via high altitude platforms. These exciting new applications will provide fertile ground for the development of new signal processing techniques.

Some of these new signal processing techniques will undoubtedly involve data-driven machine learning (ML), which is very much becoming a reality within the scope of the TC. The SPCOM area has traditionally been blessed with well-established generative models for point-to-point communication and with the existence of fundamental information theoretical limits for these models. Further, the TC has always placed high value on mathematically provable optimality of the methods that we develop. In spite of this, the adoption of data-driven methods is now moving beyond the initial

(continued on page 183)

Digital Object Identifier 10.1109/MSP.2018.2841413
Date of publication: 28 August 2018