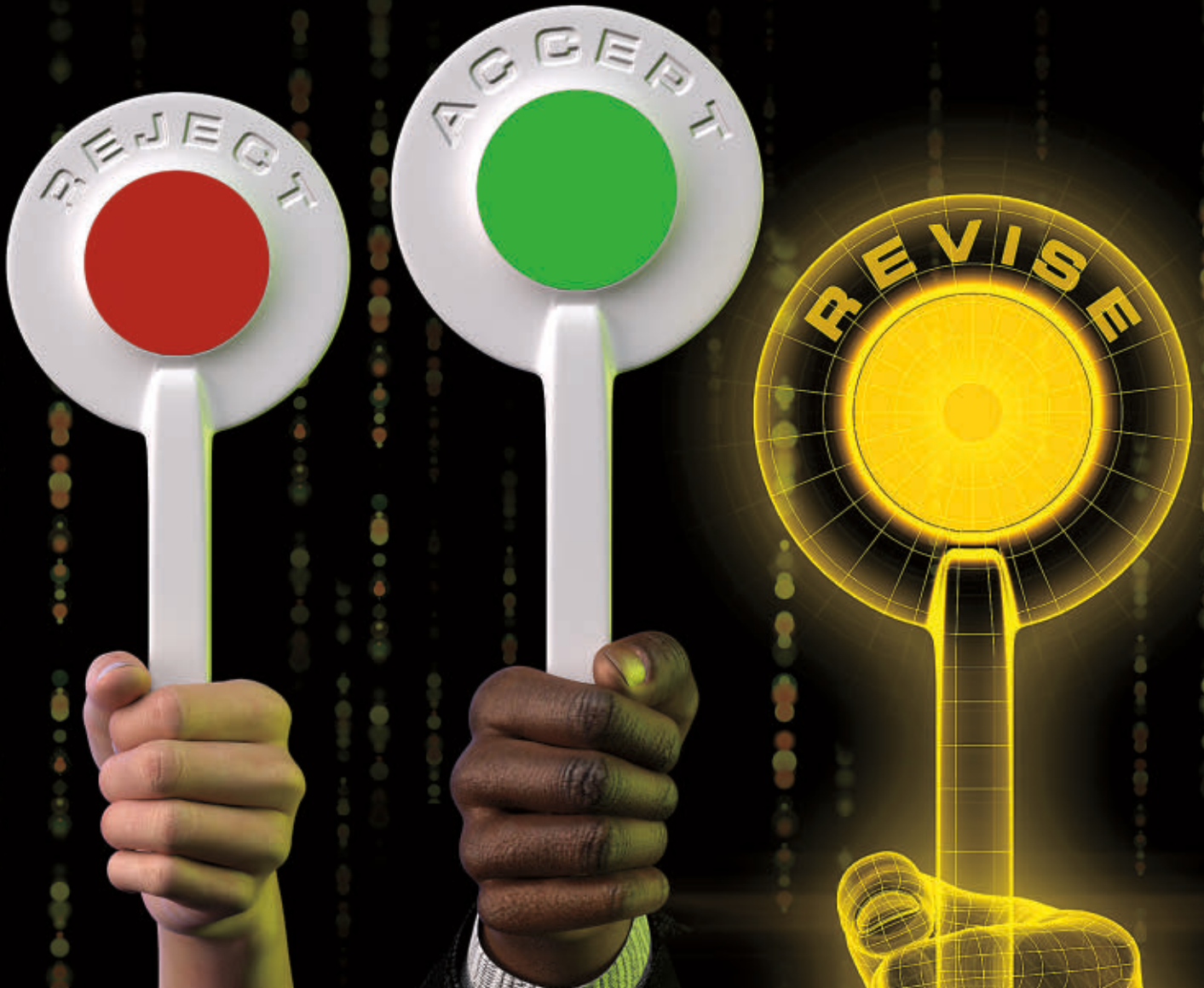


# COMMUNICATIONS

CACM.ACM.ORG OF THE ACM 03/2017 VOL.60 NO.03



## **Computational Support for Academic Peer Review A Perspective from Artificial Intelligence**

Financing the Dark Web  
ACM's Open-Conference Principle and Political Reality  
Heterogeneous Computing  
Career Histories of Top CIOs

# CELEBRATING 50 YEARS OF COMPUTING'S GREATEST ACHIEVEMENTS

Since its inauguration in 1966, the ACM A. M. Turing Award has recognized major contributions of lasting importance in computing. Through the years, it has become the most prestigious technical award in the field, often referred to as the "Nobel Prize of computing."

ACM will celebrate 50 years of the Turing Award and the visionaries who have received it with a conference on June 23 - 24, 2017 at the Westin St. Francis in San Francisco. ACM Turing laureates will join other ACM award recipients and experts in moderated panel discussions exploring how computing has evolved and where the field is headed. Topics include:

- **Advances in Deep Neural Networks**
- **Restoring Personal Privacy without Compromising National Security**
- **Moore's Law Is Really Dead: What's Next?**
- **Quantum Computing: Far Away? Around the Corner? Or Maybe Both at the Same Time?**
- **Challenges in Ethics and Computing**
- **Preserving Our Past for the Future**
- **Augmented Reality: From Gaming to Cognitive Aids and Beyond**

We hope you can join us in San Francisco, or via our live web stream, to look ahead to the future of technology and innovation, and to help inspire the next generation of computer scientists to invent and dream.

For more information and to reserve your spot, visit [www.acm.org/turing-award-50](http://www.acm.org/turing-award-50)

## Program Committee

Craig Partridge  
*Program Chair*

Fahad Dogar  
*Deputy Program Chair*

Karen Breitman

Vint Cerf

Jeff Dean

Joan Feigenbaum

Wendy Hall

Joseph Konstan

David Patterson



CELEBRATING 50 YEARS  
OF COMPUTING'S GREATEST ACHIEVEMENTS



ACM Multimedia is the premier international conference for multimedia. Here, experts and practitioners from around the world display their scientific achievements and innovative industrial products. Founded in 1993, ACM SIGMM will hold its 25<sup>th</sup> multimedia conference in Silicon Valley's renowned Computer History Museum.

We have prepared an extensive program consisting of technical sessions covering all aspects of the multimedia field. The 2017 ACM Multimedia Conference will feature oral and poster presentations, tutorials, panels, exhibits, demonstrations, and workshops. We will highlight works that bring the principal subjects of investigations into focus as well as competitions between research teams on challenging problems. Moreover, our interactive art program will stimulate artists and computer scientists to collaboratively discover the frontiers of artistic communication. Details and updates are published on the conference website:

[www.acmmm.org/2017](http://www.acmmm.org/2017).

### ACM Multimedia 2017 is welcoming a number of different contribution types:

Regular Scientific Papers	Industry Exhibitions
Brave New Idea Papers	Business Idea Venture Program
Interactive Art Exhibition	Open Source Software Competition
Multimedia Grand Challenge Competition	Workshops
Tutorials	Panels
Makers' Program	Doctoral Symposium
Technical Demonstrations	Video Program

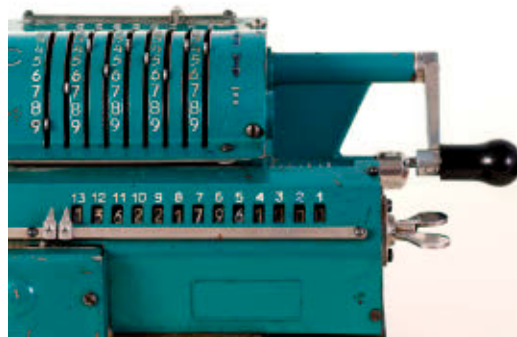
Notes: This is a partial list, please visit the website for more details. This year we have unified long and short papers into a single scientific papers submission and review process. The unified track has a flexible range of paper lengths (6-8 pages plus references), supporting shorter as well as longer papers.

### Submission Deadline:

Regular Papers Abstracts:	07 April 2017	Brave New Ideas:	31 May 2017
Regular Papers Manuscripts:	10 April 2017	Grand Challenge Solutions:	14 July 2017
Open Source Software Compet.:	28 May 2017	Interactive Artworks:	08 June 2017
Demos:	31 May 2017	Doctoral Symposium:	31 May 2017
Video Program:	16 June 2017	Panel Proposals:	31 May 2017
Workshop Proposals:	06 February 2017	Tutorial Proposals:	31 May 2017
Workshop Papers:	19 July 2017		

### Conference Location: Computer History Museum in Silicon Valley (Mountain View)

The Computer History Museum is dedicated to the preservation and celebration of computer history and is home to the largest international collection of computing artifacts in the world, encompassing computer hardware, software, documentation, ephemera, photographs, oral histories, and moving images. We will hold the 25<sup>th</sup> ACM International Conference on Multimedia in this unique location and the nearby MS Silicon Valley Campus Conference facilities.



## Departments

- 5 **Editor's Letter**  
**ACM's Open-Conference Principle and Political Reality**  
*By Moshe Y. Vardi*
- 
- 7 **From the President**  
**ACM's Commitment to Accessibility**  
*By Vicki L. Hanson*
- 
- 9 **Cerf's Up**  
**Grumpy Old Cells**  
*By Vinton G. Cerf*
- 
- 10 **Letters to the Editor**  
**Address the Consequences of AI in Advance**
- 
- 12 **BLOG@CACM**  
**The Slow Evolution of CS for All, the Beauty of Programs**  
Mark Guzdial considers the steps needed to reach the goal of CS for All, while Robin K. Hill ponders the aesthetics of programming.
- 
- 35 **Calendar**
- 
- 101 **Careers**

## Last Byte

- 104 **Q&A**  
**Out of Bounds**  
Mathematics led Subhash Khot, developer of the Unique Games Conjecture, to computer science without his ever having seen a computer.  
*By Leah Hoffmann*

## News



- 15 **Thinking Deeply to Make Better Speech**  
More work is needed to make synthesized speech more natural, easier to understand, and more pleasant to hear.  
*By Neil Savage*
- 
- 18 **The Future of Semiconductors**  
Researchers are looking for new ways to advance semiconductors as Moore's Law approaches its limits.  
*By Samuel Greengard*
- 
- 21 **Financing the Dark Web**  
Cryptocurrencies are enabling illegal or immoral transactions in the dark corners of the Internet.  
*By Keith Kirkpatrick*
- 
- 23 **ACM Recognizes New Fellows**

## Viewpoints



- 26 **Legally Speaking**  
**Supreme Court on Design Patent Damages in *Samsung v. Apple***  
Considering influences leading to the recent U.S Supreme Court decision in a years-long case that Apple filed against Samsung over iPhone design infringement.  
*By Pamela Samuelson*
- 
- 29 **Computing Ethics**  
**Where Review Goes Wrong**  
Examining professional misconduct among academic publication examiners.  
*By Elizabeth Varki*
- 
- 31 **The Profession of IT**  
**Misconceptions About Computer Science**  
Common misconceptions about computer science hinder professional growth and harm the identity of computing.  
*By Peter J. Denning, Matti Tedre, and Pat Yongpradit*
- 
- 34 **Viewpoint**  
**Learning with Mobile Technologies**  
Considering the challenges, commitments, and quandaries.  
*By Thomas M. Philip*



Practice

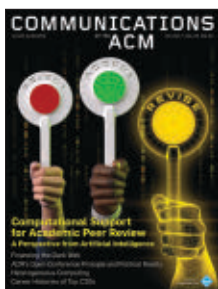


38 **Time, but Faster**  
A computing adventure about time through the looking glass.  
*By Theo Schlossnagle*

42 **Heterogeneous Computing: Here to Stay**  
Hardware and software perspectives.  
*By Mohamed Zahran*

46 **Research for Practice: Distributed Transactions and Networks as Physical Sensors**  
Expert-curated guides to the best of CS research.

**Q** Articles' development led by **acmqueue**  
[queue.acm.org](http://queue.acm.org)



**About the Cover:**  
Simon Price and Peter A. Flach examine the new tools from machine learning and AI designed to support and even automate parts of the peer-review process (p. 70). Cover illustration by Spooky Pooka at Debut Art.

Contributed Articles



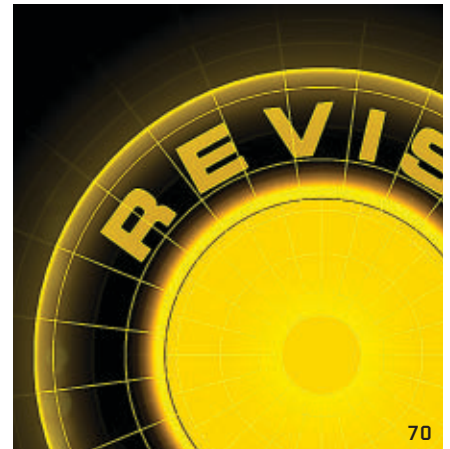
50 **Making the Field of Computing More Inclusive**  
More accessible conferences, digital resources, and ACM SIGs will lead to greater participation by more people with disabilities.  
*By Jonathan Lazar, Elizabeth F. Churchill, Tovi Grossman, Gerrit Van der Veer, Philippe Palanque, John "Scooter" Morris, and Jennifer Mankoff*



Watch the authors discuss their work in this exclusive *Communications* video.  
<http://cacm.acm.org/videos/making-the-field-of-computing-more-inclusive>

60 **The Path to the Top: Insights from Career Histories of Top CIOs**  
Along the way, acquire technical expertise and a master's degree, even while changing positions and companies.  
*By Daniel J. Mazzola, Robert D. St. Louis, and Mohan R. Tanniru*

Review Articles



70 **Computational Support for Academic Peer Review: A Perspective from Artificial Intelligence**  
New tools tackle an age-old practice.  
*By Simon Price and Peter A. Flach*



Watch the authors discuss their work in this exclusive *Communications* video.  
<http://cacm.acm.org/videos/computational-support-for-academic-peer-review>

Research Highlights

- 82 **Technical Perspective**  
**The Power of Wi-Fi to Deliver Power**  
*By Srinivasan Keshav*
- 83 **Powering the Next Billion Devices with Wi-Fi**  
*By Vamsi Talla, Bryce Kellogg, Benjamin Ransford, Saman Naderiparizi, Joshua R. Smith, and Shyamnath Gollakota*
- 92 **Technical Perspective**  
**Data Distribution for Fast Joins**  
*By Leonid Libkin*
- 93 **Reasoning on Data Partitioning for Single-Round Multi-Join Evaluation in Massively Parallel Systems**  
*By Tom J. Ameloot, Gaetano Geck, Bas Ketsman, Frank Neven, and Thomas Schwentick*



ACM, the world's largest educational and scientific computing society, delivers resources that advance computing as a science and profession. ACM provides the computing field's premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources.

**Executive Director and CEO**

Bobby Schnabel  
**Deputy Executive Director and COO**  
 Patricia Ryan

**Director, Office of Information Systems**  
 Wayne Graves

**Director, Office of Financial Services**  
 Darren Ramdin

**Director, Office of SIG Services**  
 Donna Cappel

**Director, Office of Publications**  
 Scott E. Delman

**ACM COUNCIL**

**President**

Vicki L. Hanson

**Vice-President**

Cherri M. Pancake

**Secretary/Treasurer**

Elizabeth Churchill

**Past President**

Alexander L. Wolf

**Chair, SGB Board**

Jeanna Matthews

**Co-Chairs, Publications Board**

Jack Davidson and Joseph Konstan

**Members-at-Large**

Gabriele Anderst-Kotis; Susan Dumais; Elizabeth D. Mynatt; Pamela Samuelson; Eugene H. Spafford

**SGB Council Representatives**

Paul Beame; Jenna Neefe Matthews; Barbara Boucher Owens

**BOARD CHAIRS**

**Education Board**

Mehran Sahami and Jane Chu Prey

**Practitioners Board**

Terry Coatta and Stephen Ibaraki

**REGIONAL COUNCIL CHAIRS**

**ACM Europe Council**

Dame Professor Wendy Hall

**ACM India Council**

Srinivas Padmanabhuni

**ACM China Council**

Jianguang Sun

**PUBLICATIONS BOARD**

**Co-Chairs**

Jack Davidson; Joseph Konstan

**Board Members**

Ronald F. Boisvert; Karin K. Breitman; Terry J. Coatta; Anne Condon; Nikil Dutt; Roch Guerin; Carol Hutchins; Yannis Ioannidis; Catherine McGeoch; M. Tamer Ozsu; Mary Lou Soffa; Alex Wade; Keith Webster

**ACM U.S. Public Policy Office**

Renee Dopplick, Director  
 1701 Pennsylvania Ave NW, Suite 300,  
 Washington, DC 20006 USA  
 T (202) 659-9711; F (202) 667-1066

**Computer Science Teachers Association**

Mark R. Nelson, Executive Director

# COMMUNICATIONS OF THE ACM

Trusted insights for computing's leading professionals.

*Communications of the ACM* is the leading monthly print and online magazine for the computing and information technology fields. *Communications* is recognized as the most trusted and knowledgeable source of industry information for today's computing professional. *Communications* brings its readership in-depth coverage of emerging areas of computer science, new trends in information technology, and practical applications. Industry leaders use *Communications* as a platform to present and debate various technology implications, public policies, engineering challenges, and market trends. The prestige and unmatched reputation that *Communications of the ACM* enjoys today is built upon a 50-year commitment to high-quality editorial content and a steadfast dedication to advancing the arts, sciences, and applications of information technology.

**STAFF**

**DIRECTOR OF PUBLICATIONS**

Scott E. Delman  
 cacm-publisher@cacm.acm.org

**Executive Editor**

Diane Crawford

**Managing Editor**

Thomas E. Lambert

**Senior Editor**

Andrew Rosenbloom

**Senior Editor/News**

Larry Fisher

**Web Editor**

David Roman

**Rights and Permissions**

Deborah Cotton

**Art Director**

Andrij Borys

**Associate Art Director**

Margaret Gray

**Assistant Art Director**

Mia Angelica Balaquiot

**Designer**

Iwona Usakiewicz

**Production Manager**

Lynn D'Addesio

**Advertising Sales Account Manager**

Ilia Rodriguez

**Columnists**

David Anderson; Phillip G. Armour; Michael Cusumano; Peter J. Denning; Mark Guzdial; Thomas Haigh; Leah Hoffmann; Mari Sako; Pamela Samuelson; Marshall Van Alstyne

**CONTACT POINTS**

**Copyright permission**  
 permissions@hq.acm.org

**Calendar items**  
 calendar@cacm.acm.org

**Change of address**  
 acmhhelp@acm.org

**Letters to the Editor**  
 letters@cacm.acm.org

**WEBSITE**

http://cacm.acm.org

**AUTHOR GUIDELINES**

http://cacm.acm.org/

**ACM ADVERTISING DEPARTMENT**

2 Penn Plaza, Suite 701, New York, NY  
 10121-0701  
 T (212) 626-0686  
 F (212) 869-0481

**Advertising Sales Account Manager**

Ilia Rodriguez  
 ilia.rodriguez@hq.acm.org

**For display, corporate/brand advertising:**

Craig Pitcher  
 pitcherc@acm.org T (408) 778-0300

**Media Kit** acmm mediasales@acm.org

**Association for Computing Machinery (ACM)**

2 Penn Plaza, Suite 701  
 New York, NY 10121-0701 USA  
 T (212) 869-7440; F (212) 869-0481

**EDITORIAL BOARD**

**EDITOR-IN-CHIEF**

Moshe Y. Vardi  
 eic@cacm.acm.org

**NEWS**

**Co-Chairs**

William Pullleyblank and Marc Snir

**Board Members**

Mei Kobayashi; Michael Mitzenmacher; Rajeev Rastogi; François Sillion

**VIEWPOINTS**

**Co-Chairs**

Tim Finin; Susanne E. Hambrusch; John Leslie King

**Board Members**

William Aspray; Stefan Bechtold; Michael L. Best; Judith Bishop; Stuart I. Feldman; Peter Freeman; Mark Guzdial; Rachelle Hollander; Richard Ladner; Carl Landwehr; Carlos Jose Pereira de Lucena; Beng Chin Ooi; Loren Terveen; Marshall Van Alstyne; Jeannette Wing

**PRACTICE**

**Co-Chair**

Stephen Bourne

**Board Members**

Eric Allman; Peter Bailis; Terry Coatta; Stuart Feldman; Benjamin Fried; Pat Hanrahan; Tom Killalea; Tom Limoncelli; Kate Matsudaira; Marshall Kirk McKusick; George Neville-Neil; Theo Schlossnagle; Jim Waldo

The Practice section of the CACM Editorial Board also serves as the Editorial Board of *COMMUNIQUE*.

**CONTRIBUTED ARTICLES**

**Co-Chairs**

Andrew Chien and James Larus

**Board Members**

William Aiello; Robert Austin; Elisa Bertino; Gilles Brassard; Kim Bruce; Alan Bundy; Peter Buneman; Peter Druschel; Carlo Ghezzi; Carl Gutwin; Yannis Ioannidis; Gal A. Kaminka; James Larus; Igor Markov; Gail C. Murphy; Bernhard Nebel; Lionel M. Ni; Kenton O'Hara; Sriram Rajamani; Marie-Christine Rousset; Avi Rubin; Krishan Sabnani; Ron Shamir; Yoav Shoham; Larry Snyder; Michael Vitale; Wolfgang Wahlster; Hannes Werthner; Reinhard Wilhelm

**RESEARCH HIGHLIGHTS**

**Co-Chairs**

Azer Bestavros and Gregory Morrisett

**Board Members**

Martin Abadi; Amr El Abbadi; Sanjeev Arora; Michael Backes; Nina Balcan; Andrei Broder; Doug Burger; Stuart K. Card; Jeff Chase; Jon Crowcroft; Alexei Efros; Alon Halevy; Norm Jouppi; Andrew B. Kahng; Sven Koenig; Xavier Leroy; Steve Marschner; Kobbi Nissim; Guy Steele, Jr.; Margaret H. Wright; Nikolai Zeldovich; Andreas Zeller

**WEB**

**Chair**

James Landay

**Board Members**

Marti Hearst; Jason I. Hong; Jeff Johnson; Wendy E. MacKay

**ACM Copyright Notice**

Copyright © 2017 by Association for Computing Machinery, Inc. (ACM). Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and full citation on the first page. Copyright for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or fee. Request permission to publish from permissions@hq.acm.org or fax (212) 869-0481.

For other copying of articles that carry a code at the bottom of the first or last page or screen display, copying is permitted provided that the per-copy fee indicated in the code is paid through the Copyright Clearance Center; www.copyright.com.

**Subscriptions**

An annual subscription cost is included in ACM member dues of \$99 (\$40 of which is allocated to a subscription to *Communications*); for students, cost is included in \$42 dues (\$20 of which is allocated to a *Communications* subscription). A nonmember annual subscription is \$269.

**ACM Media Advertising Policy**

*Communications of the ACM* and other ACM Media publications accept advertising in both print and electronic formats. All advertising in ACM Media publications is at the discretion of ACM and is intended to provide financial support for the various activities and services for ACM members. Current advertising rates can be found by visiting http://www.acm-media.org or by contacting ACM Media Sales at (212) 626-0686.

**Single Copies**

Single copies of *Communications of the ACM* are available for purchase. Please contact acmhhelp@acm.org.

**COMMUNICATIONS OF THE ACM**

(ISSN 0001-0782) is published monthly by ACM Media, 2 Penn Plaza, Suite 701, New York, NY 10121-0701. Periodicals postage paid at New York, NY 10001, and other mailing offices.

**POSTMASTER**

Please send address changes to *Communications of the ACM*  
 2 Penn Plaza, Suite 701  
 New York, NY 10121-0701 USA

Printed in the U.S.A.



Association for Computing Machinery





Moshe Y. Vardi

DOI:10.1145/3047270

# ACM's Open-Conference Principle and Political Reality

ACM's Open-Conference Statement starts with a lofty principle: "The open exchange of ideas and the freedom of thought and expression are central to the aims and goals of ACM

and its conferences. These aims and goals require an environment that recognizes the inherent worth of every person and group, that fosters dignity, understanding, and mutual respect, and that embraces diversity." (See <https://www.acm.org/conferences>.) This principle reflects ACM's mission of "advance computing as a science and a profession; enable professional development; and promote policies and research that benefit society." In the past few weeks, however, this principle has been gravely tested.

In March 2016, the U.S. State of North Carolina passed a sweeping law (House Bill 2—HB2) that reversed a local ordinance that had extended some rights to people who are gay or transgender. The new law also nullified local ordinances around the state that would have expanded protections for the LGBT community. Several U.S. localities issued travel bans in response to HB2, limiting travel to North Carolina. In January 2017, the ACM SIGMOD Executive Committee decided to move the SIGMOD/PODS 2017 conference out of North Carolina to a new location (see statement here: <http://wp.sigmod.org/?p=2079>). This decision resolved the issue for one conference. Unfortunately, in the 2017 legislative session, state legislators in 11 other U.S. states have pre-filed or introduced legislation that would restrict access to multiuser restrooms, locker rooms, and other sex-segregated facilities on the basis of a certain definition of sex or gender ("bathroom bills"). The SIGMOD/PODS 2018 conference is cur-

rently slated for Houston, Texas, but this plan is now in jeopardy as Texas is one of the states that is discussing passing a "bathroom bill."

But the bathroom-bill issue was dwarfed by an Executive Order issued by U.S. President Trump on Jan. 27, 2017, which banned nationals of seven Muslim-majority countries from entering the U.S. for at least the next 90 days. This includes persons with valid U.S. visas, as well as—at least initially—U.S. permanent residents. This Executive Order covers not only new arrivals to the U.S., but also persons who have been residing in the U.S. and are temporarily outside the U.S. In response to this executive order, ACM expressed grave concerns and urged the lifting of the visa suspension so as not to curtail the studies or contributions of scientists and researchers. I'd like to see ACM go farther and band with other professional societies to fight the Executive Order; perhaps this will have happened by the time this letter is published.

As this issue goes to print, we do not know how the status of the Executive Order will unfold. There are strong arguments against the constitutionality of the Order, and lawsuits against the U.S. government have already been filed. But it may take months if not years, for the legal process to conclude, and the outcome is far from certain. In the meantime, if we follow the SIGMOD precedent, ACM should avoid holding conferences in the U.S. Should it? I think not.

In fact, while I appreciate the reasoning that led the SIGMOD Executive Committee to decide to relocate the 2017 conference away from North Carolina, I disagree with the decision. ACM is a global professional society. Its Open-Conference Principle has to be interpreted from that perspective. Undoubtedly, there are going to be more liberal and less liberal interpretations. Should all ACM conferences be held in California, which tends to be the most liberal state in the U.S.? Or, in view of the Executive Order, how about moving all ACM conferences to Sweden? This is not only impractical, but, in my opinion, not even right. In fact, the ACM SIGMOD/PODS 2007 conference was held in Beijing, China. There are those who would have argued then that China's human-rights record is not up to Western standards, so ACM should not hold conferences in China. But the ACM SIGMOD Executive Committee decided then, correctly, I believe, that going to China rather than avoiding China would better serve the case of open conferences.

The Open-Conference Principle is aimed at benefiting society. When we boycott a particular locality, we are also telling our colleagues in that locality, who are likely to be supporting the cause of open and just society, that we would rather stay away than come and support their fight. This is unlikely, I believe, to benefit society. Boycotting may feel right, but I doubt that it would be productive. Staying and fighting for a cause—though it is far from clear what the best way of doing it is—may be much harder, but is the right path.

Follow me on Facebook, Google+, and Twitter.

**Moshe Y. Vardi**, EDITOR-IN-CHIEF

Copyright held by author.



ACM Books



MORGAN & CLAYPOOL  
PUBLISHERS

# Publish your next book in the ACM Digital Library

ACM Books is a new series of advanced level books for the computer science community, published by ACM in collaboration with Morgan & Claypool Publishers.

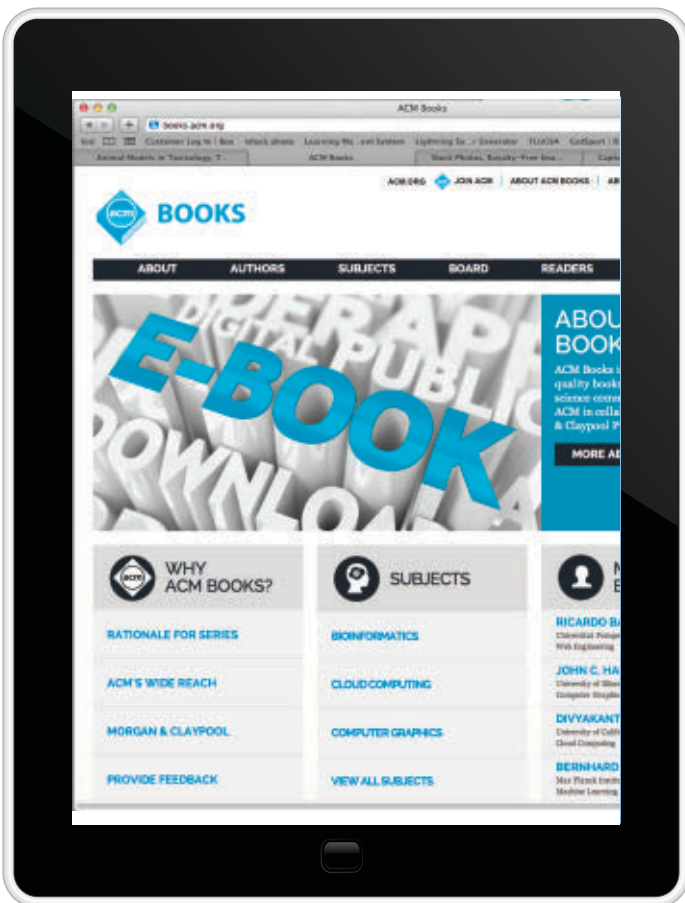
*I'm pleased that ACM Books is directed by a volunteer organization headed by a dynamic, informed, energetic, visionary Editor-in-Chief (Tamer Özsu), working closely with a forward-looking publisher (Morgan and Claypool).*

—Richard Snodgrass, University of Arizona

[books.acm.org](http://books.acm.org)

## ACM Books

- ◆ will include books from across the entire spectrum of computer science subject matter and will appeal to computing practitioners, researchers, educators, and students.
- ◆ will publish graduate level texts; research monographs/overviews of established and emerging fields; practitioner-level professional books; and books devoted to the history and social impact of computing.
- ◆ will be quickly and attractively published as ebooks and print volumes at affordable prices, and widely distributed in both print and digital formats through booksellers and to libraries and individual ACM members via the ACM Digital Library platform.
- ◆ is led by EIC M. Tamer Özsu, University of Waterloo, and a distinguished editorial board representing most areas of CS.



**Proposals and inquiries welcome!**

Contact: **M. Tamer Özsu**, Editor in Chief  
[booksubmissions@acm.org](mailto:booksubmissions@acm.org)



Association for  
Computing Machinery

*Advancing Computing as a Science & Profession*





Vicki L. Hanson

DOI:10.1145/3047268

# ACM's Commitment to Accessibility

It is no secret that my passion for being an ACM volunteer began with SIGACCESS—the ACM Special Interest Group on Accessibility and Computing. As a new volunteer, I was

highly motivated by a talk by Ben Shneiderman in which he said he was proud to be part of an organization that had, as part of its code of ethics, the following: *“In a fair society, all individuals would have equal opportunity to participate in, or benefit from, the use of computer resources regardless of race, sex, religion, age, disability, national origin, or other such similar factors.”* I, too, am proud to be part of a society that supports these goals.

As ACM's President, I remain focused on issues of diversity. I would like to highlight two key aspects of accessibility already being addressed by ACM. The first is digital accessibility; the second is conference accessibility.

**Digital accessibility.** PDFs in the ACM Digital Library typically are not accessible. Within the next year, however, a new set of conference and journal templates will be rolled out that will include enhanced accessibility features. Working with a new publications vendor, automatic accessibility features will be created wherever possible and features that require author input will be flagged for author attention as part of the production process. Both accessible PDFs and HTML5 will be the outputs and will begin populating the Digital Library. Of course, as with all new processes, we can expect a few bumps along the way. Before release, however, ACM will have thoroughly tested these enhanced documents with users representative of disability communities.

Notably, ACM has committed to subtitling/closed captioning all video

materials released by ACM. Thus, all new content on the ACM YouTube channel will be accessible. This will benefit not only those with a hearing loss, but also should prove helpful to individuals who do not have English as their first language.

The new ACM website, which rolled out last year, had accessibility as an explicit requirement. Representatives from SIGACCESS and SIGCHI were involved in shaping the details of these requirements. Critically, users from disability communities were involved in testing pages generated by the new ACM page template.

An **ACM Web Accessibility Statement** is included on the ACM website, <http://www.acm.org/accessibility>. It is worth mentioning that much material on that website comes from vendors and volunteers, not from ACM headquarters. ACM is working with these contributors to make their offerings accessible. The **ACM Web Accessibility Statement** includes an *accessibility style guide* to assist those contributors in making their content consistent with ACM's Web accessibility standards.

**As ACM's President, I remain focused on issues of diversity.**

**Conference accessibility.** ACM has long had a commitment to making conferences accessible for attendees as well as presenters. SIGACCESS and SIGCHI again are leaders in conference accessibility (For details, see the article by Lazar et al. on page 50). The SIGACCESS ASSETS conference has long been a valuable proving ground given the high proportion of ASSETS attendees experiencing disability. The SIGACCESS conference guidelines provide a useful source of information for other organizers on how to create an accessible conference, <http://www.sigaccess.org/welcome-to-sigaccess/resources/accessible-conference-guide>.

For conferences that occur within the U.S., venues should meet ADA (Americans with Disabilities Act) requirements. Conferences that occur outside the U.S. are subject to local regulations governing accessibility. Each conference's contact at ACM's SIG Services is knowledgeable and committed to providing conference experiences that are accessible. Conference organizers are encouraged to avail themselves of these services. In all cases, ACM strives to meet the needs of attendees who need accommodation. This is a long-standing commitment and an aspect of ACM that sets it apart as a premier professional society and about which we can all be rightfully proud. □

Vicki L. Hanson (vlh@acm.org) is ACM President, Distinguished Professor at Rochester Institute of Technology, and a professor at the University of Dundee. Twitter: @ACM\_President.

Copyright held by author.



## Seeking applications from outstanding young leaders

The ACM Future of Computing Academy will bring together next-generation researchers, practitioners, educators and entrepreneurs from various disciplines of computing to define and launch new initiatives that will carry us into the future. Academy members will have the satisfaction of contributing to our field while enjoying the opportunity to grow their personal networks across regions, computing disciplines and computing professions. ACM invites accomplished professionals typically in their 20s and 30s to apply.

The inaugural meeting of the ACM Future of Computing Academy will be June 25, 2017 in San Francisco. Members of the Academy will be invited to attend ACM's celebration of 50 Years of the ACM Turing Award, June 23-24, at the Westin St. Francis, where they will have the opportunity to interact with ACM A.M. Turing Award laureates.



"Academy members will have the privilege and responsibility of being the voice of the next generation of computing professionals and ensuring that ACM continues to contribute to their success long into the future."

– **Vicki Hanson**, ACM President

"The Future of Computing Academy will give some of the most talented, creative, and passionate young computing professionals a collective voice that will help shape the future of our industry and its influence on our social and economic ecosystem."

– **Vint Cerf**, Google Chief Internet Evangelist and former ACM President



"The Future of Computing Academy will afford members an invaluable opportunity to expand their professional networks to include outstanding individuals with demonstrated excellence from across a breadth of computing disciplines."

– **Aaron Quigley**, Chair of Human Computer Interaction, University of St. Andrews

"Members of the Academy will have opportunities to interact with computing pioneers whose foundational contributions influence innovation today."

– **Matthias Kaiserswerth**, Managing Director, Hasler Foundation



Apply at: <http://www.acm.org/fca>



Association for  
Computing Machinery



Vinton G. Cerf

DOI:10.1145/3028774

## Grumpy Old Cells

**I** AM GOING way out on a limb in this column into an area where I really know very little but am completely fascinated by what I am learning. The tenuous linkage to our discipline is what I will call *programmed cell self-destruction* or maybe *cell suicide*.

I have been reading at length a book called *Molecular Biology of the Cell*. This is a 1,342-page book, not counting index and glossary and a separate book of problems. It is profusely illustrated and eminently readable even by a layperson like me, pretending, of course, that I am actually understanding what I am reading.

It turns out that cells reproduce (that is, divide: *mitosis*) but usually only a finite number of times. When they divide, their DNA is duplicated within the cell and separate copies are transported into each new daughter cell. Human DNA comes in 23 distinct chromosomes. Each chromosome is made up of a double helix of DNA. During cell mitosis, each strand of the double helix is duplicated by figuratively *unzipping* the double helix and replicating each strand. The replication takes place at multiple replication origin sites along the strand so this process operates in parallel. The now-duplicated chromosomes look like elongated “X”-shaped Gumby characters formed by adjacent DNA strands. As the mitotic process continues, the duplicated chromosomes are pulled apart by microtubules that attach to opposite sides of the paired chromosomes. As the process proceeds, eventually two new nuclei form with its copy of the original cell’s DNA and the cell completes its division into two essentially identical cells.

At the ends of each strand of DNA is a repetitive sequence of DNA called a *telomere*. There are multiple

telomeres at each end of the chromosome. One might think of them figuratively as handles needed to anchor the DNA during the unzipping and replication process. The telomeres themselves are not replicated in this process, so every cell division may lose one or more telomeres. If there are too few telomeres left, the replication process fails and initiates a process known as cell *apoptosis*, which we can think of as programmed cell death. Interestingly, certain kinds of cells known as embryonic stem (ES) cells found in bone marrow and in the gut contain an enzyme called *telomerase* that fabricates new telomeres so that ES cells can replicate indefinitely. Non-ES cells, which make up most cells in our bodies, count down to termination of replication. Cancer cells manage to avoid this outcome by using telomerase to make more telomeres allowing indefinite proliferation.

Cellular apoptosis can be triggered within the cell or by outside factors

**No matter how complicated we think our software systems have become, we can still marvel at the extraordinary complexity of the life of a single cell.**

and the disintegrated debris, each enclosed in cell membranes, are consumed by *phagocytes* (“cell eaters”). It is estimated, for example, that phagocytes consume  $10^{11}$  blood cells a day. Other cells experience *necrosis*, which is triggered from outside the cell causing the cell to rupture and spew its contents into intercellular space. One way this can happen is if cells have experienced some form of trauma. Viruses can invade cells, commandeer the DNA interpretation system (that is, *ribosomes*) with their own DNA, and eventually rupture the cell wall, broadcasting new virus particles into the surrounding tissues.

Once a cell has stopped replicating, it may not immediately experience either apoptosis or necrosis. Rather, it may continue to exist in a senescent state, which I have chosen to label an increasingly grumpy state. It may continue to produce proteins but they may prove to be harmful to other cells. The aging process and its manifest side effects can be traced, in part, to grumpy old cells spewing harmful products into the biological neighborhood. One thinks of the *plaques* and *tangles* of Alzheimer’s disease and the misfolded *prions* associated with Creutzfeldt-Jakob encephalopathy caused by harmful proteins synthesized by grumpy cells.

If you got all the way to the end of this column, congratulations! No matter how complicated we think our software systems have become, we can still marvel at the extraordinary complexity of the life of a single cell and the immeasurable complexity of multicellular life, including our own. **□**

Vinton G. Cerf is vice president and Chief Internet Evangelist at Google. He served as ACM president from 2012–2014.

Copyright held by author.

# Address the Consequences of AI in Advance

**T**HE VIEWPOINTS by Alan Bundy “Smart Machines Are Not a Threat to Humanity” and Devdatt Dubhashi and Shalom Lappin “AI Dangers: Imagined and Real” (both Feb. 2017) argued against the possibility of a near-term singularity wherein super-intelligent AIs exceed human capabilities and control. Both relied heavily on the lack of direct relevance of Moore’s Law, noting raw computing power does not by itself lead to human-like intelligence. Bundy also emphasized the difference between a computer’s efficiency in working an algorithm to solve a narrow, well-defined problem and human-like generalized problem-solving ability. Dubhashi and Lappin noted incremental progress in machine learning or better knowledge of a biological brain’s wiring do not automatically lead to the “unanticipated spurts” of progress that characterize scientific breakthroughs.

These points are valid, but a more accurate characterization of the situation is that computer science may well be just one conceptual breakthrough away from being able to build an artificial general intelligence. The considerable progress already made in computing power, sensors, robotics, algorithms, and knowledge about biological systems will be brought to bear quickly once the architecture of “human-like” general intelligence is articulated. Will that be tomorrow or in 10 years? No one knows. But unless there is something about the architecture of human intelligence that is ultimately inaccessible to science, that architecture will be discovered. Study of the consequences is not premature.

**Martin Smith**, McLean, VA

## ACM Code of Ethics vs. Autonomous Weapons

“Can We Trust Autonomous Weapons?” as Keith Kirkpatrick asked at the top of his news story (Dec. 2016).

Autonomous weapons already exist on the battlefield (we call them land mines and IEDs), and, despite the 1997 Ottawa Mine Ban Treaty, we see no decrease in their use. Moreover, the decision as to whether to use them is unlikely to be left to those who adhere to the ACM Code of Ethics. The Washington Naval Treaty of 1922 was concluded between nation-states—entities that could be dealt with in historically recognized ways, including sanctions, demarches, and wars. An international treaty between these same entities regarding autonomous weapons would have no effect on groups like ISIS, Al-Qaida, Hezbollah, the Taliban, or Boko Haram. Let us not be naïve ... They have access to the technology, knowledge, and materials to create autonomous weapons, along with the willingness to use them. When they do, the civilized nations of the world will have to decide whether to respond in kind—defensive systems with sub-second response times—or permit their armed forces to be outclassed on the battlefield. I suspect the decision will seem obvious to them at the time.

**Joseph M. Saur**, Virginia Beach, VA

It was rather jarring to read in the same issue (Dec. 2016) a column “Making a Positive Impact: Updating the ACM Code of Ethics” by Bo Brinkman et al. on revamping the Code and a news article “Can We Trust Autonomous Weapons?” by Keith Kirkpatrick on autonomous weapons. Such weapons are, of course, enabled entirely by software that is presumably written by at least some ACM members. How does the Code’s “Do no harm” ideal align with building devices whose sole reason for existing is to inflict harm? It seems that unless this disconnect is resolved the Code is aspirational at best and in reality a generally ignored shelf-filling placeholder.

**Jack Ganssle**, Reisterstown, MD

## To Model Complexity in Fiction, Try Fractals

Robin K. Hill raised an interesting point in her blog post “Fiction as Model Theory” (Dec. 2016) that fictional characters and worlds need to follow certain rules—rules that can be formalized and verified for consistency. Fiction in general, and science fiction in particular, has always been of considerable interest to scholarly researchers. What was notable in Hill’s post was her suggestion of using formalism in rather unconventional domains—domains not traditionally identified with computation-related methods.

I have personally taken a similar path and, together with my colleagues, discovered the utility of formalizing ideas from unconventional domains. These range from modeling complex living environments in self-organizing arrays of motion sensors to identifying unexpected emergent patterns in the spread of disease in large-scale human populations or even in cousin marriages.<sup>1</sup> Likewise, I have found that formal specification can prove useful in terms of representing community-identified cognitive development of scholarly researchers measured as a function of their citation indices.<sup>2</sup>

Could a longer work of fiction, say, a novel or novella, benefit from such treatment? After all, well-written novels often invent their own internally consistent landscapes. They also often involve a rather complex interplay of characters, multiple plotlines, backstories, and conflicts. Scholarly researchers have even identified social networks of fictional characters influencing major events in these make-believe worlds. It is indeed the interplay of characters in conflict that makes for a potential page-turner or, at least, a novel worth reading.

While fiction authors have developed their own instruments, ranging from Randy Ingermanson’s so-called “snowflake method” to Shawn Coyne’s

“story grid” for editors, what is of particular interest to me is the recurrence of self-similar patterns in well-written fiction. Snowflakes consist of fractals, and Coyne has identified similar patterns in well-written novels repeating in sub-scenes he calls “beats” and in scenes, scene sequences, and even the Aristotelian three-act structure; that is, same pattern, different scales. The “story grid” method performs a quantitative dissection of fiction, allowing editors to help create generally engaging fiction.

Fractals, or mathematical sets repeating at multiple scales, appear frequently in nature. Examples range from Romanesco broccoli to river basins and ferns. Prominent identification of fractal-related scholarly work includes the Mandelbrot set, Serpinski’s carpet, Koch Snowflake, Julia set, strange attractor, and unified mass central triangle. We can thus infer well-written works of fiction might be better modeled through a *combination* of formal specification and fractals. Formalism could thus be useful even for people associated with the novel-publishing industry.

#### References

1. Akhtar, N., Niazi, M., Mustafa, F., and Hussain, A. A discrete event system specification (DEVS)-based model of consanguinity. *Journal of Theoretical Biology* 285, 1 (Sept. 2011), 103–112.
2. Hussain, A. and Niazi, M. Toward a formal, visual framework of emergent cognitive development of scholars. *Cognitive Computation* 6, 1 (Mar. 2014), 113–124.

**Muaz A. Niazi**, Islamabad, Pakistan

#### No Hologram from HoloLens

Although Marina Krakovsky’s news article “Bringing Holography to Light” (Oct. 2016) was timely (the visual interface will indeed dominate the future), the photo in the article’s Figure 1 above the caption “Learning medicine in three dimensions with Microsoft’s HoloLens.” was completely opposite of what Krakovsky said in the article’s opening sentence. Microsoft HoloLens is not even designed to produce a holographic image. On the contrary, Microsoft HoloLens is just a see-through stereoscopic head-mounted display, with two diffractive mirrors that are prefabricated diffractive reflection lenses manufactured either by dia-

mond turning or optical holography. There is neither holographic processing nor holographic image reconstruction. In the HoloLens, a stereoscopic image pair is projected before the user’s eyes through the diffractive mirrors. There is a marked difference between a stereoscopic 3D image and a holographic image. A holographic image can reproduce true 3D perspectives, whereas a stereoscopic 3D image cannot.

**Debesh Choudhury**, Kolkata,  
West Bengal, India

#### Why Not Trisexuality?

Adi Livnat and Christos Papadimitriou review article “Sex as an Algorithm” (Nov. 2016) was fascinating but mis-titled. It discussed the benefits of conjugality. George C. Williams in *Sex and Evolution* distinguished the more general concept conjugality from (eu)sexuality, in which the number of conjugal strains in the species is equal to the number of individuals participating in conjugation—two, in all conjugal species on this planet. This seems an important distinction, and I suggest the cover of *Communications* was misleading. In my own book *Albatross I* emphasized this and other distinctions, aiming to avoid nonsensical talk, as in that arising from “the gostak distims the doshes” in *The Meaning of Meaning* by C.K. Ogden and I.A. Richards.

Livnat’s and Papadimitriou’s reference to their non-coverage of heterozygosity was revealing. I rather suspect heterozygosity is a prerequisite for sexuality proper; certainly a lot of sexual species are haploid in the gametic generation and diploid in the others.

Some of the mathematics as to the binarity of conjugation might be interesting. What are the chances that on some other world there may have evolved life with a triple helix, ternary conjugation—and so trisexuality?

**John A. Wills**, Oakland, CA

*Communications* welcomes your opinion. To submit a Letter to the Editor, please limit yourself to 500 words or less, and send to [letters@cacm.acm.org](mailto:letters@cacm.acm.org).

© 2017 ACM 0001-0782/17/3 \$15.00

Coming Next Month in **COMMUNICATIONS**

**Attack of the  
Killer Microseconds**

**A Service  
Computing Manifesto**

**Computational Thinking  
for Teacher Education**

**Guilt-Free Data Reuse**

**Wanted: Toolsmiths**

**Kickstarter’s  
Yancey Strickler on  
Today’s Entrepreneur**

**Uninitialized Reads**

**Pervasive, Dynamic  
Authentication of  
Physical Items**

**Certifying a File System  
Using Crash Hoare Logic**

**Does Anyone  
Listen to You?**

Plus the latest news about implantable wireless sensors, mapping technologies, and computing the arts.

The *Communications* Web site, <http://cacm.acm.org>, features more than a dozen bloggers in the BLOG@CACM community. In each issue of *Communications*, we'll publish selected posts or excerpts.



Follow us on Twitter at <http://twitter.com/blogCACM>

DOI:10.1145/3037383

<http://cacm.acm.org/blogs/blog-cacm>

## The Slow Evolution of CS for All, the Beauty of Programs

*Mark Guzdial considers the steps needed to reach the goal of CS for All, while Robin K. Hill ponders the aesthetics of programming.*



### Mark Guzdial Taking Incremental Steps Toward CS for All

<http://bit.ly/2gCFpSM>  
November 28, 2016

At the end of October, the Expanding Computing Education Pathways (ECEP) alliance organized a summit with the White House Office of Science and Technology Policy (OSTP) on state implementation of the President's CS for All initiative. You can see the agenda at <http://bit.ly/2ifPVwY> and a press release on the two days of meetings at <http://bit.ly/2iMvyeK>. I learned a lot at those meetings; one insight I gained was that the CS for All initiative will succeed in increments. U.S. states are developing novel, incremental approaches to CS for All.

The event's second day was focused on teams from the 16 states and Puerto Rico in the ECEP Alliance. At a session on teacher certifications, some of the attendees were concerned with what they saw as lowering standards in order to get more certified teachers. "We have a shortage of doctors in rural areas. That

doesn't mean we make it easier to become a doctor!" That made sense to me, but then I heard others push the metaphor a bit. Adding more nurses and more physician assistants does improve quality of care, and it is less expensive to have more of these health care providers than to produce enough doctors.

Only a few U.S. states offer CS teacher initial certification, which requires a choice to become a CS teacher while still an undergraduate and take years of classes. Georgia and California, like several other states, offer an add-on certification ("endorsement") teachers can earn after gaining a certification in something else. An endorsement typically still requires multiple semester-long courses. Utah has one of the most innovative CS teacher add-on certification schemes, with three levels: an initial level that requires only some summer professional development, and two further levels requiring post-secondary courses.

Leigh Ann DeLyser hosted a great session about CSNYC and the new CS for All Consortium. CSNYC is charged with

implementing Mayor Bill de Blasio's initiative to make CS education available to all students in all grades in all New York City schools by 2025. DeLyser told us CS-NYC is defining the Mayor's initiative as a school-based mandate. Even 10 years and \$81 million isn't enough to provide certified, full-time CS teachers in every school so every student gets a CS course.

Rather, every school must offer to every student in every grade a high-quality CS learning experience. Maybe that's a full course, like the BJC CS Principles curriculum now in NYC schools. Alternatively, it might be a Bootstrap unit in an algebra class, or a CT STEM activity that uses StarLogo to achieve NGSS science learning goals. It's a reasonable incremental approach towards CS for All.

New Hampshire, one of the newest ECEP states, is exploring micro-certifications. Rather than getting a certification as a CS teacher, a mathematics or science teacher might get a micro-certification to demonstrate proficiency in using a computer science approach in their teaching. There might be micro-certificates in Bootstrap, CT STEM, or Project GUTS for middle school science.

We want a future where computer science is taught by certified teachers and is as universally available as mathematics and science classes are today in most U.S. high schools. That's the vision Briana Morrison and I wrote about in *CACM* (<http://bit.ly/2iIFeEc>). Along the way, we need ways of growing CS education where we develop teachers who know about and teach computer science, even if not full-time, certified CS teachers.



## Robin K. Hill What Makes a Program Elegant?

<http://bit.ly/2e2U6yK>

October 11, 2016

A subfield of philosophy is aesthetics, in which we attempt to understand beauty. Is beauty universal? Does it make us better people somehow? Why do we focus on beauty, not ugliness? A ready application of this question to computer science (CS) addresses program elegance. Most programmers, or so I believe, would agree some programs are elegant, and elegant programs are better than others, and experienced programmers, or so I believe, generally agree on which programs are elegant.

The criterion of efficiency looms large in production programming, and appears in comment on elegance on the Web, for instance by Perrin (<http://bit.ly/2ih2lhR>). A program should be brief, but not a slave to brevity. An elegant design artifact is sleek and spare in its utility. An elegant program is minimally gratuitous. Consider Binary Search (of an ordered sequence) as opposed to Sequential Search, or Quicksort as opposed to Insertion Sort (<http://bit.ly/2j71dcx>). Sequential Search tediously examines each (ordered) item, but does not have to; Bubble Sort tediously exchanges many items that will have to be moved again. To find the first  $n$  prime numbers, we can tediously test each for divisors or we can deploy the Sieve of Eratosthenes. Efficiency helps make the Sieve, Binary Search, and Quicksort elegant. We have our first criterion for elegance, **(1) minimality**, encompassing both shortness and simplicity.

Let's avoid features of programs depending on source code syntax, or compilers, or I/O mechanisms, or memory handling. A program that minimizes temporary variables, directly evaluating expressions instead, is "better," but we do not address the question of aesthetics at that level, nor at the level of self-describing identifiers, nor documentation, nor modularity, nor design patterns. A program also becomes better as it includes more error-checking, which does not strengthen, and may weaken, its elegance even as it enhances its quality.

Simplicity by itself can't be enough; Bubblesort is a simple program. (I would count Boyer-Moore String Search as el-

egant, though it's complicated.) Brevity by itself can't be enough; the C loop control `while(i++ < 10)` is terse, excelling in brevity, but its elegance is debatable. I would call it, in the architectural sense, brutalism. Architecture provides nice analogues because it also strives to construct artifacts that meet specifications under material constraints, prizing especially those artifacts that manifest beauty as well (<http://bit.ly/2j8AMkN>).

A factor that looms larger in CS than in architecture or other disciplines is correctness. A building may be regarded as elegant even if marginal parts of it are uncomfortable, but no program that does not work is regarded as elegant. This gives us another criterion, **(2) accomplishment**—the program does what it is supposed to do. Though included in the list of desiderata here, failure on that criterion is fatal rather than detrimental.

Constraints under which programming is done impose a context without which the elegance cannot be appreciated. We must understand the problem, the tools, and materials, to appreciate the solution. Expertise is necessary. Examining many student programs over many years refines an appreciation ever more impressed by work that does it all with graceful assurance and economy. Elegance, therefore, is doubly relative—to the context of the work and to the background of the observer.

Bitmap Sort, as presented by Jon Bentley (<http://bit.ly/2ikzqSE>) in a classic "Programming Pearls" column, is still worth studying. To sort  $n$  unique integers in a fixed range  $0$  to  $m$ , we rearrange them through a comparison-based sort such as Quicksort, or we initialize a bit array, indexed by  $0$  to  $m$ , to *false*, and then for each integer input, flip its bit to *true*. A pass through the resulting array, during which the indices of the *true* bits are output, gives us the sorted list. This is nice, and elegant, even relative to Quicksort, but only works on a set of unique values (as described); recognition of situations that meet that restriction distinguishes the programmer of elegance.

We are ducking hard questions about implementations at various levels of translation, and whether they should count toward or against elegance, and we will continue to do so. In fact, what I have been describing is not programs in source code terms, but algorithms. Brevity, or minimality, is a salient fea-

ture of code, but a subtle feature of algorithms; what we want is minimality in terms of the solution, however that solution is expressed. Yet another more general concept of spareness is at play in elegance, something like restraint. This gives us a criterion of **(3) modesty**. An example that flouts it comes right off the very first page of another classic, Kernighan and Plauger's *Elements of Programming Style* (<http://bit.ly/2ikHDq8>):

```
DO 14 I=1,N      DO 14 J=1,N  14 V(I,J)
= (I/J)*(J/I)
```

This exploits the FORTRAN compiler's truncation of integer division results to populate a matrix  $V$  with zeroes everywhere except the diagonal, where the values are one; that is, it initializes  $V$  to the  $N \times N$  identity matrix. This is clever and short, but oh, dear, it's implementation-dependent, therefore fragile; it's obscure and ostentatious. Such virtuosity is unfortunate, yet hard to resist. (Kernighan and Plauger propose the obvious initialization to zero throughout, followed by a loop that assigns the value one to each  $V(N,N)$ .)

What else counts? An elegant program confers a sense of satisfaction, of enlightenment. Let's call this criterion, especially characteristic of program artifacts, **(4) revelation**—the program shows us something new about its task, or brings to the fore something we forgot. Eratosthenes' Sieve shows us, or reminds us, multiples are the "not-primes." Bitmap Sort shows us, or reminds us, the integers are already ordered; they come as a sequence, so sorting can be accomplished by an indication of presence only. Boyer-Moore String Search shows us strings are just as distinct backward as they are forward.

The criteria for program elegance suggested here are **(1) minimality**, **(2) accomplishment**, **(3) modesty**, and **(4) revelation**, all rooted in the particulars of the problem. Are these criteria necessary? Sufficient? Inadequate? Because of dependence on the problem at hand, sometimes with complex circumstances, a wide range of examples of elegant programs is difficult to come by. What exemplars stand out in your world? ■

Mark Guzdial is a professor at the Georgia Institute of Technology. Robin K. Hill is an adjunct professor at the University of Wyoming.

© 2017 ACM 0001-0782/17/3 \$15.00

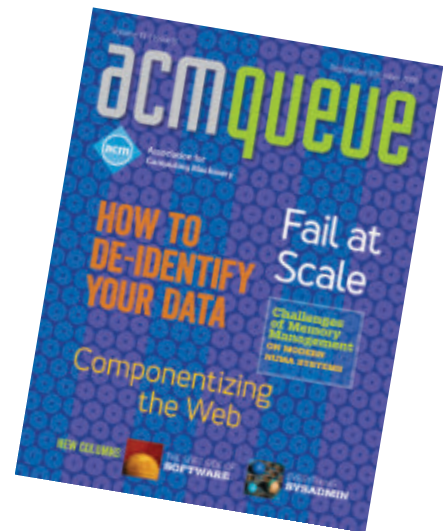
# acmqueue

Check out the new acmqueue app

FREE TO ACM MEMBERS

acmqueue is ACM's magazine by and for practitioners, bridging the gap between academics and practitioners of computer science. After more than a decade of providing unique perspectives on how current and emerging technologies are being applied in the field, the new acmqueue has evolved into an interactive, socially networked, electronic magazine.

Broaden your knowledge with technical articles focusing on today's problems affecting CS in practice, video interviews, roundtables, case studies, and lively columns.



Keep up with this fast-paced world on the go. Download the mobile app.



Association for  
Computing Machinery

Desktop digital edition also available at [queue.acm.org](http://queue.acm.org).  
Bimonthly issues free to ACM Professional Members.  
Annual subscription \$19.99 for nonmembers.



## Thinking Deeply to Make Better Speech

*More work is needed to make synthesized speech more natural, easier to understand, and more pleasant to hear.*

**M**ACHINES THAT SPEAK are nothing new. Siri has been answering questions from iPhone users since 2011, and text-to-voice programs have been around even longer. People with speaking disabilities—most famously, Stephen Hawking—have used computers to generate speech for decades. Yet synthesizing speech that sounds as natural as if spoken by a human is still an elusive goal, although one that appears to be getting closer to reality.

If you listen to the latest version of Apple's Siri, "it sounds pretty amazing," says Simon King, a professor of speech processing and director of the Centre for Speech Technology Research at the University of Edinburgh. Apple, Google, and Microsoft all have commercial speech applications that read text in a neutral but reasonable-sounding tone. Words are pronounced correctly, for the most part, and generally flow from one to the next in perfectly acceptable sentences. "We're quite good at that and the speech is very intelligible," King says.

Researchers in speech synthesis, however, would like to move beyond merely "intelligible" to speech that



**A humanoid robot, named Aiko Chihira by its creators at Toshiba and Osaka University, at a 2015 trial in Tokyo's Mitsukoshi department store. Toshiba says it will incorporate speech recognition and synthesis into the robot by 2020.**

sounds more natural. Their work could make synthesized speech easier to understand and more pleasant to hear. It could also allow them to synthesize better voices for people unable to speak for themselves, and create text-to-speech systems for less-common languages.

"Practically all the systems work well at the sentence level," says Alex Acero, senior director of Siri at Apple. Ask a machine to read you a newspaper article or an email message from your mother, however, and the result will be flat. "Yes, you can understand it if you pay attention, but it's still not the same

as having someone read it to you,” he says. Computerized speech cannot handle prosody—the rhythm and intonation of speech that conveys meaning and adds emotional context. “That is incredibly important for humans,” says Acero. “That’s why when you send text messages, you add emojis.”

There are two basic approaches to creating speech. The older one is parametric speech synthesis, in which a computer generates sounds from the elements of text. Over the years, that has evolved into statistical parametric speech synthesis, which uses a statistical model to create the proper waveform for each sound. For a long time the statistical model used was a hidden Markov model, which calculates the future state of a system based on its current state. In the past couple of years, however, hidden Markov models been replaced with deep neural networks, which compute the interaction between different factors in successive layers. That switch, King says, has led to an improvement in the accuracy of the parametric approach.

The technique that has mainly been used over the last couple of decades is concatenative speech synthesis, in which a human speaker records many hours of speech, which is then diced into individual units of sound called phonemes and then spliced back together to create new phrases that the original speaker never uttered. Apple, for instance, splits the phonemes, represented as waveforms, in half. That provides more choices for finding different phonemes that fit together smoothly, Acero explains.

The latest iteration of Siri combines parametric and concatenative speech synthesis. It relies on a statistical model called a mixed density network—a type of neural network—to learn the parameters of the phonemes it is looking for, examining hundreds of features such as whether a sound is stressed or not, or which phonemes usually proceed or follow others. Once it knows what the waveforms of the speech are supposed to look like, it searches for appropriate ones in the recorded speech and fits them together. The system does not necessarily create every phrase from scratch; groups of words and sometimes even whole sen-

## As good as Siri is, its speech lacks prosody—the rhythm and intonation of speech that conveys meaning and adds emotional context.

tences can be taken directly from the recording. “It is more automated and it’s more accurate because it’s more data-driven,” says Acero.

As good as the results are, however, the speech still lacks prosody, because the machine does not really understand what it is saying. That lack may explain one problem with synthesized speech, King believes; while it may be completely intelligible to someone in a quiet room who is paying attention, if the listener is in a noisy environment, or trying to multitask, or has hearing loss or dyslexia, the intelligibility drops off much more rapidly than it does with natural speech.

King hypothesizes the drop-off occurs because natural speech contains a lot of redundancies, cues that aid in understanding what is being said. There may be, for instance, changes in intonation or stress or pitch when one word leads into another in natural speech. Such acoustic cues are not there in synthesized speech, and in concatenative speech words plucked from different sentences may even contain the wrong cues.

It may also be that having to process such inconsistencies makes the listener’s brain work harder, which may increase the chances of missing something. “You couldn’t say your synthetic speech is truly natural until it’s as good as natural speech for everybody in every environment,” King says.

“In order to say something in the most natural way, you pretty much need to understand what it means,” King says. Though speech recognition is good enough for Siri and

similar systems to respond to questions and commands, their level of understanding is still fairly shallow, he says. They can recognize individual words, identify nouns and verbs, notice local sentence structure, even distinguish a question from a statement. Researchers working on natural language understanding are using approaches such as vector spaces, which focus on statistics such as how frequently words appear, but so far machines are not able to understand speech—especially in large chunks such as paragraphs or entire passages—on a deep-enough level to be able to read them the way a human would.

### A New Wave

Last September, Google announced it had made great strides with a technique called WaveNet. Developed by DeepMind, a London-based company that Google bought in 2014, WaveNet uses statistical parametric synthesis relying on deep neural networks to produce speech in both English and Mandarin that listeners rated as superior to the best existing systems (there is no objective measurement of speech quality, so it is always assessed by human listeners). The system also automatically generated piano music. Google published its results in a blog post and in a paper on ArXiv, but declined to make the researchers available for press interviews.

Google’s approach was inspired by a model it had published earlier in the year that used a neural network to generate natural-looking images one pixel at a time. The researchers trained the system by feeding it waveforms recorded from human speakers. Such raw audio can contain 16,000 samples per second, so it is computationally expensive. Once trained, they fed the system text they had broken down into a sequence of linguistic and phonetic features, giving the computer such information as what word, syllable, and phoneme it was seeing. They were able to train it on different speakers so it could speak in different voices, and provided it with different accents and emotions.

Acero calls WaveNet a very interesting approach, which somewhere down the road might replace concatenative

synthesis. At the moment, though, it takes several hours of computing to produce one second of speech, so it is not immediately practical.

### A Physical Model


Oriol Guasch, a physicist and mathematician at Ramon Llull University in Barcelona, Spain, is also taking a computationally intensive approach to speech synthesis. He is working on mathematically modeling the entire human vocal tract. “We’d like to simulate the whole physical process, which will, in the end, generate the final sound,” he says.

To do that, he takes an MRI image of a person’s vocal tract as he is pronouncing, say, the vowel “E.” He then represents that geometry of the vocal folds, soft palate, lips, nose, and other parts with differential equations. Using that, he generates a computational mesh, a many-sided grid that approximates the geometry. The process is not easy; a desktop computer can generate a mesh with three to four million elements in about three or four hours to represent the short “A” sound, he says. A sibilant “S,” though, requires a computer with 1,000 processors to run for a week to generate 45 million elements. The added complexity of that sound arises from the air flowing between the teeth and creating turbulent eddies swirling in complex patterns. Imagine, then, the time required to produce a whole word, let alone a sentence.

Guasch sees his approach more as an interesting computing challenge than a practical attempt to create speech. “The final goal is not just synthesizing speech, it’s about reproducing the way the human body behaves,” he says. “I believe when you have a computational problem, it’s good to face it from many different angles.”

The University of Edinburgh’s King, on the other hand, is working toward practical applications. He recently received funding for a three-year project, in conjunction with the BBC World Service, to create text-to-speech systems for languages that do not have enough speakers to make developing a system a financially attractive process for companies. It should be possible to use machine learning

on data such as radio broadcasts and newspapers to build a credible system, King says, without the expense of hiring linguistic experts and professional voice artists. He has already built a Swahili prototype, which he says works pretty well.

King also has developed a system can take a small number of recordings of a particular individual’s speech and apply them to a model already trained with a much larger dataset, and use that to generate new speech that sounds like that individual. The system is undergoing clinical trials in a U.K. hospital to see if it can be a practical way of helping people with amyotrophic lateral sclerosis, who are expected to lose their ability to speak as their disease progresses. “This is not going to help them live any longer, but for the time they do live it could help make their quality of life better,” he says. 

### Further Reading

*Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., and Kalchbrenner, N.*  
WaveNet: A Generative Model for Raw Audio, ArXiv, Cornell University Library, 2016  
<http://arxiv.org/pdf/1609.03499>

*King, S., and Karaiskos, V.*  
The Blizzard Challenge 2016, Blizzard Challenge Workshop, Sept. 2016, Cupertino, CA  
[http://www.festvox.org/blizzard/bc2016/blizzard2016\\_overview\\_paper.pdf](http://www.festvox.org/blizzard/bc2016/blizzard2016_overview_paper.pdf)

*Arnela, M., Dabbaghchian, S., Blandin, R., Guasch, O., Engwall, O., Van Hirtum, A., and Pelorson, X.*

Influence of vocal tract geometry simplifications on the numerical simulation of vowel sounds, *Journal of the Acoustical Society of America*, 140, 2016  
<http://dx.doi.org/10.1121/1.4962488>

*Deng, L., Li, J., Huang, J-T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., Gong, Y., and Acero, A.*

Recent advances in deep learning for speech research at Microsoft, IEEE International Conference on Acoustics, Speech and Signal Processing, 2013  
[http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=6639345](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6639345)

Simon King – Using Speech Synthesis to Give Everyone Their Own Voice  
<https://www.youtube.com/watch?v=xzL-pxcpe-E>

Neil Savage is a science and technology writer based in Lowell, MA.

© 2017 ACM 0001-0782/17/3 \$15.00

# ACM Member News

## TRYING TO DETERMINE WHAT IS COMPUTABLE



**Santosh Vempala,**  
Distinguished  
Professor of  
Computer  
Science in the  
College of

Computing at the Georgia Institute of Technology (Georgia Tech), earned his undergraduate degree in computer science at the Indian Institute of Technology, New Delhi, India, and his Ph.D. in Algorithms, Combinatorics, and Optimization from Carnegie Mellon University in 1997. “What I really wanted to study was theory of computation; what is computable and what is not, with what amount of resources.”

After obtaining his doctorate, Vempala became a professor of mathematics at the Massachusetts Institute of Technology, a post he held for almost 10 years before moving to Georgia Tech in 2006. There, Vempala served as the first director (from 2006 to 2011) of the Algorithms and Randomness Center, a think tank dedicated to exploring the theory of computing and optimization.

His research focuses on the intersection of algorithms, randomness, and geometry. “The relationship between algorithms and geometry has been mutually beneficial,” Vempala explains. Initially it was about using techniques mathematicians had developed to work with algorithms in new ways, but now the questions and answers have made a deep contribution to the development of algorithmic geometry, he says.

Vempala continues to be fascinated by whether certain problems have efficient solutions. Some of his recent research has been focused on trying to understand how the brain works, and the modeling of its computational abilities.

In 2008, Vempala launched an initiative called Computing for Good, which develops deployable computing solutions for social problems like inequality, homelessness, and healthcare delivery, in areas where resources are constrained.

—John Delaney

# The Future of Semiconductors

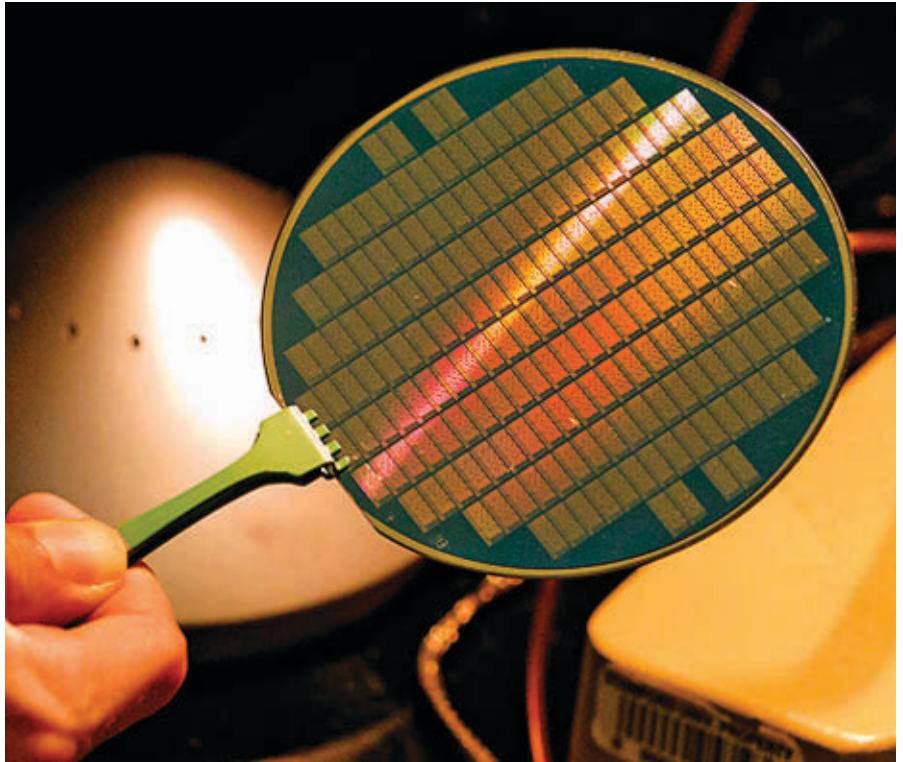
*Researchers are looking for new ways to advance semiconductors as Moore's Law approaches its limits.*

**O**VER THE LAST half-century, as computing has advanced by leaps and bounds, one thing has remained fairly static: Moore's Law. The concept, named after semiconductor pioneer Gordon Moore, is based on the observation that the number of transistors packed into an integrated circuit (IC) doubles approximately every two years. For more than 50 years, this concept has provided a predictable framework for semiconductor development. It has helped computer manufacturers and many other companies focus their research and plan for the future.

However, there are signs that Moore's Law is reaching the end of its practical path. Although the IC industry will continue to produce smaller and faster transistors over the next few years, these systems cannot operate at optimal frequencies due to heat dissipation issues. This has "brought the rate of progress in computing performance to a snail's pace," wrote IEEE fellows Thomas M. Conte and Paolo A. Gargini in a 2015 IEEE-RC-ITRS report, *On the Foundation of the New Computing Industry Beyond 2020*.

Yet, the challenges do not stop there. There is also the fact that researchers cannot continually miniaturize chip designs; at some point over the next several years, current two-dimensional ICs will reach a practical size limit. Although researchers are experimenting with new materials and designs—some radically different—there currently is no clear path to progress. In 2015, Gordon Moore predicted the law that bears his name will wither within a decade. The IEEE-RC-ITRS report noted: "A new way of computing is urgently needed."

As a result, the semiconductor industry is in a state of flux. There is a growing recognition that research



**Engineers at Stanford University are creating wafers like this one from carbon nanotubes, a potential successor to silicon that could make processors smaller and more energy efficient.**

and development must incorporate new circuitry designs and rely on entirely different methods to scale up computing power further. "For many years, engineers didn't have to work all that hard to scale up performance and functionality," observes Jan Rabaey, professor and EE Division Chair in the Electrical Engineering and Computer Sciences Department at the University of California, Berkeley. "As we reach physical limitations with current technologies, things are about to get a lot more difficult."

## The Incredible Shrinking Transistor

The history of semiconductors and Moore's Law follows a long and somewhat meandering path. Conte, a professor at the schools of computer science

and engineering at Georgia Institute of Technology, points out that computing has not always been tied to shrinking transistors. "The phenomenon is only about three decades old," he points out. Prior to the 1970s, high-performance computers, such as the CRAY-1, were built using discrete emitter-coupled logic-based components. "It wasn't really until the mid-1980s that the performance and cost of microprocessors started to eclipse these technologies," he notes.

At that point, engineers developing high-performance systems began to gravitate toward Moore's Law and adopt a focus on microprocessors. However, the big returns did not last long. By the mid-1990s, "The delays in the wires on-chip outpaced the delays due to transis-

tor speeds,” Conte explains. This created a “wire-delay wall” that engineers circumvented by using parallelism behind the scenes. Simply put: the technology extracted and executed instructions in parallel, but independent, groups. This was known as the “super-scalar era,” and the Intel Pentium Pro microprocessor, while not the first system to use this method, demonstrated the success of this approach.

Around the mid-2000s, engineers hit a power wall. Because the power in CMOS transistors is proportional to the operating frequency, when the power density reached 200W/cm<sup>2</sup>, cooling became imperative. “You can cool the system, but the cost of cooling something hotter than 150 watts resembles a step function, because 150 watts is about the limit for relatively inexpensive forced-air cooling technology,” Conte explains. The bottom line? Energy consumption and performance would not scale in the same way. “We had been hiding the problem from programmers. But now we couldn’t do that with CMOS,” he adds.

No longer could engineers pack more transistors onto a wafer with the same gains. This eventually led to reducing the frequency of the processor core and introducing multicore processors. Still, the problem didn’t go away. As transistors became smaller—hitting approximately 65nm in 2006—performance and economic gains continued to subside, and as nodes dropped to 22nm and 14nm, the problem grew worse.

What is more, all of this has contributed to fabrication facilities becoming incredibly expensive to build, and semiconductors becoming far more expensive to manufacture. Today, there are only four major semiconductor manufacturers globally: Intel, TSMC, GlobalFoundries, and Samsung. That is down from nearly two dozen two decades ago.

To be sure, the semiconductor industry is approaching the physical limitations of CMOS transistors. Although alternative technologies are now in the research and development stage—including carbon nanotubes and tunneling field effect transistors (TFETs)—there is no evidence these next-gen technologies will actually pay off in a major way. Even if they do usher

**“There are technical and engineering challenges, economic challenges because we’re seeing fewer industry players, and fundamental changes in the way people use computing devices.”**

in further performance gains, they can at best stretch Moore’s Law by a generation or two.

In fact, industry groups such as the IEEE International Roadmap of Devices and Systems (IRDS) initiative have reported it will be nearly impossible to shrink transistors further by 2023.

Observes Michael Chudzik, a senior director at Applied Materials: “Semiconductor technology is challenged on many fronts. There are technical and engineering challenges, economic challenges because we’re seeing fewer industry players, and fundamental changes in the way people use computing devices” such as smartphones, as well as cloud computing, and the Internet of Things (IoT), which place entire different demands on ICs. This makes the methods of the past less desirable in the future. “We are entering a different era,” Rabaey observes.

### Designs on the Future

Mapping out a future for integrated circuits and computing is paramount. One option for advancing chip performance is the use of different materials, Chudzik says. For instance, researchers are experimenting with cobalt to replace tungsten and copper in order to increase the volume of the wires, and studying alternative materials for silicon. These include Ge, SiGE and III-V materials such as gallium arsenide and gallium indium arsenide. However, these materials present performance and scaling challenges and, even if those problems can be addressed, they would produce only incre-

mental gains that would tap out in the not-too-distant future.

Faced with the end of Moore’s Law, researchers are also focusing attention on new and sometimes entirely different approaches. One of the most promising options is stacking components and scaling from today’s 2D ICs to 3D designs, possibly by using nanowires. “By moving into the third dimension and stacking memory and logic, we can create far more function per unit volume,” Rabaey explains. Yet, for now, 3D chip designs also run into challenges, particularly in terms of cooling. The devices have less surface volume as engineers stack components. As a result, “You suddenly have to do processing at a lower temperature or you damage the lower layers,” he notes.

Consequently, a layered 3D design, at least for now, requires a fundamentally different architecture. “Suddenly, in order to gain denser connectivity, the traditional approach of having the memory and processor separated doesn’t make sense. You have to rethink the way you do computation,” Rabaey explains. It’s not an entirely abstract proposition. “The advantages that some applications tap into—particularly machine learning and deep learning, which require dense integration of memory and logic—go away.” Adding to the challenge: a 3D design increases the risk of failures within the chip. “Producing a chip that functions with 100% integrity is impossible. The system must be fail-tolerant and deal with errors,” he adds.

Regardless of the approach and the combination of technologies, researchers are ultimately left with no perfect option. Barring a radical breakthrough, they must rethink the fundamental way in which computing and processing take place.

Conte says two possibilities exist beyond pursuing the current technology direction.

One is to make radical changes, but limit these changes to those that happen “under the covers” in the microarchitecture. In a sense, this is what took place in 1995, except “today we need to use more radical approaches,” he says. For servers and high-performance computing, for example, ultra-low-temperature superconducting is being advanced as one possible solution. At present, the U.S. Intelligence Advanced

Research Projects Activity (IARPA) is investing heavily in this approach within its Cryogenic Computing Complexity (C3) program. These non-traditional logic gates are made in small scale, at a size roughly 200 times larger than today's transistors.

Another is to “bite the bullet and change the programming model,” Conte says. Although numerous ideas and concepts have been forwarded, most center on creating fixed-function (non-programmable) accelerators for critical parts of important programs. “The advantage is that when you remove programmability, you eliminate all the energy consumed in fetching and decoding instructions.” Another possibility—and one that is already taking shape—is to move computation away from the CPU and toward the actual data. Essentially, memory-centric architectures, which are in development in the lab, could muscle up processing without any improvements in chips.

Finally, researchers are exploring completely different ways to compute, including neuromorphic and quantum models that rely on non-Von-Neumann brain-inspired methods and quantum computing. Rabaey says processors are already heading in this direction. As

deep learning and cognitive computing emerge, GPU stacks are increasingly used to accelerate performance at the same or lower energy cost as traditional CPUs. Likewise, mobile chips and the Internet of Things bring entirely different processing requirements into play. “In some cases, this changes the paradigm to lower processing requirements on the system but having devices everywhere. We may see billions or trillions of devices that integrate computation and communication with sensing, analytics, and other tasks.”

In fact, as visual processing, big data analytics, cryptography, AR/VR, and other advanced technologies evolve, it is likely researchers will marry various approaches to produce boutique chips that best fit the particular device and situation. Concludes Conte: “The future is rooted in diversity and building devices to meet the needs of the computer architectures that have the most promise.”

#### Further Reading

Conte, T.M., and Gargini, P.A. *On the Foundation of the New Computing Industry Beyond 2020*, Executive Summary, *IEEE Rebooting Computing and ITRS*. September 2015. <http://rebootingcomputing.ieee.org/images/files/pdf/prelim-ieee-rc-itrs.pdf>

Lam, C.H. *Neuromorphic Semiconductor Memory*, 3D Systems Integration Conference (3DIC), 2015 International, 31 Aug.-2 Sept. 2015. <http://ieeexplore.ieee.org/document/7334566/>

Claeys, C., Chiappe, D., Collaert, N., Mitard, J., Radu, I., Rooyackers, R., Simoen, E., Vandooren, A., Veloso, A., Waldron, N.H. Witters, L., and Thean, A. *Advanced Semiconductor Devices for Future CMOS Technologies*, *ECS Transactions*, 66 (5) 49-60 (2015) 10.1149/06605.0049ecst ©The Electrochemical Society. 2015. [https://www.researchgate.net/profile/C\\_Claeys/publication/277896307\\_Invited\\_Advanced\\_Semiconductor\\_Devices\\_for\\_Future\\_CMOS\\_Technologies/links/565ad44408aefe619b240bcc.pdf](https://www.researchgate.net/profile/C_Claeys/publication/277896307_Invited_Advanced_Semiconductor_Devices_for_Future_CMOS_Technologies/links/565ad44408aefe619b240bcc.pdf)

Cheong, H. *Management of Technology Strategies Required for Major Semiconductor Manufacturer to Survive in Future Market*, Graduate School of Management of Technology, Hoseo University, Asan 336-795, Korea, *Procedia Computer Science* 91 (2016) 1116 – 1118. *Information Technology and Quantitative Management (ITQM 2016)*. <http://www.sciencedirect.com/science/article/pii/S1877050916313564>

Samuel Greengard is an author and journalist based in West Linn, OR.

© 2017 ACM 0001-0782/17/3 \$15.00

## Milestones

# Computer Science Awards, Appointments

### RENAMED “ACM PRIZE IN COMPUTING” TO RECOGNIZE CONTRIBUTIONS BY YOUNG PROFESSIONALS

ACM recently announced that the ACM-Infosys Foundation Award in the Computing Sciences has been renamed the ACM Prize in Computing. Infosys will continue to fund the award, which recognizes computing professionals in the early to middle stages of their careers. In conjunction with the renaming of the award, the corresponding cash prize has been increased to \$250,000.

The ACM Prize in Computing recognizes computing professionals for early-to-mid-career, fundamental, innovative contributions in computing that, through depth, impact, and broad implications, exemplify the greatest achievements in

the discipline. The inaugural ACM-Infosys Foundation Award in the Computing Sciences was awarded in 2007 to Daphne Koller.

In addition to Koller, past recipients have included Stefan Savage (2015), Dan Boneh (2014), David Blei (2013), Jeff Dean and Sanjay Ghemawat (2012), Sanjeev Arora (2011), Frans Kaashoek (2010), Eric Brewer (2009), and Jon Kleinberg (2008).

“Many people know that ACM bestows the A.M. Turing Award, often referred to as ‘the Nobel Prize of Computing’ and our field’s most prestigious honor,” explained ACM President Vicki L. Hanson. “However, by focusing on early- and mid-career professionals, the ACM Prize highlights innovations that are changing paradigms and reshaping technology in ways

that will lay future foundations in the field.”

“An awards program serves to educate the public about how important research and achievement impacts society,” adds Vishal Sikka, CEO of Infosys. “The computing field, where the pace of change is more rapid than other disciplines, has experienced unprecedented transformations during the past 10 years. In addition to giving credit to these young visionaries, the ACM Prize will enlighten the public about the underpinnings that make technological advances possible.”

Underscoring the renaming and prestige of the award, the Heidelberg Laureate Forum Foundation announced that ACM Prize in Computing recipients will now be invited to participate in the Heidelberg

Laureate Forum (HLF), an annual networking event for mathematicians and computing scientists from all over the world. Each September, HLF brings the laureates of the major awards in computer science and mathematics together with brilliant young researchers from around the globe to Heidelberg, Germany, for a week of intensive exchange. ACM Prize recipients will join laureates of the ACM A.M. Turing Award (computer science), the Abel Prize (mathematics), the Fields Medal (mathematics), and the Nevanlinna Prize (mathematics).

The recipient of the 2016 ACM Prize in Computing will be announced in April, and will be formally recognized at ACM’s annual awards banquet in San Francisco.

# Financing the Dark Web

*Cryptocurrencies are enabling illegal or immoral transactions in the dark corners of the Internet.*

**L**AW ENFORCEMENT HAS long acted in accordance with the old adage of “following the money” when trying to track down those who commit crimes. Finding out who has paid for what usually provides a pretty strong picture of a crime and its relevant actors, even if no one had specifically witnessed the actions taking place.

Yet in cyberspace, following the money can be significantly more difficult, particularly on ‘Dark Web’ sites, where any number of illegal or immoral transactions are taking place, such as the sale of drugs, prostitution and human trafficking, illegal pornography, and other unsavory activities. The Dark Web, which is a huge set of web pages that are not indexed by traditional services such as Google and require a specific browser to access, have long played host to online marketplaces offering sex, drugs, and other illegal material. These marketplaces route communications and transactions via multiple computers and layers of encryption to protect the identities of vendors and purchasers, and often use cryptocurrency to further obfuscate the identities of the transacting parties.

Indeed, bitcoin, Monera, Shadow Money, and other cryptocurrencies use encryption techniques to regulate the generation of units of currency and verify the transfer of funds, all while operating independently of a central bank. All transactions are captured on a shared, visible, and distributed ledger known as a blockchain, but the cryptographic keys and digital wallets used to hold funds are not linked to real-world identities, and provided that precautions are taken, offer a high degree of anonymity compared with traditional Western digital payment methods.

For the casual observer and law enforcement professionals, it is this anonymity that has cast a pall over bitcoin and other cryptocurrencies. “There is some stigma associated with crypto-



currencies, because it was associated with things like Silk Road,” says David Decary-Hetu, an adjunct professor of criminology at the University of Montreal, and a bitcoin enthusiast. Indeed, Silk Road (and its descendants, Silk Road II and Silk Road III) capitalized on the use of bitcoins, which further helped to obscure the identities of those purchasing drugs and other illegal paraphernalia on the platform.

“There is no way to tie your identity to your online bitcoin wallet address, if you do it properly,” Decary-Hetu says, noting cryptocurrency users that try to convert those funds to traditional money may lose that anonymity. “That’s where sloppy people are going to get arrested. If they use Coinbase or another major exchange to convert bitcoin to U.S. dollars, the user must send in a scan of your passport or identification papers. If you just sell something on a cryptomarket, and then try to convert your bitcoin to local currency, then the FBI will be able to identify you very easily.”

## Dark Web Stigma

Many people automatically associate bitcoin with the Dark Web, due to the publicity surrounding the Silk Road investigation, but there are legitimate reasons for using cryptocurrencies, according to Decary-Hetu. In particular, bitcoin is viewed as a more efficient currency to use when conducting cross-border transactions, since there are

no currency fluctuations or exchange rates with which to deal. Further, in some parts of the world, cryptocurrencies may be more efficient to use or more stable than government-backed currencies. Moreover, some individuals simply may want the anonymity to purchase items that may not be illegal, but perhaps embarrassing.

“There are many helpful and legal reasons for having bitcoin,” Decary-Hetu says, noting that large established companies such as Dell Computer, Expedia, Microsoft, and PayPal, each accept bitcoin, and are clearly not dealing in illegal goods.

“Cryptocurrencies are not illegal per se,” Decary-Hetu says. “Are they helping money laundering? Probably at some level, but it might be too harsh to say that they’re only for illegal purchases.”

Still, all hope of tracking down and identifying bad actors that use cryptocurrencies is not lost, though most of the information retrieved by law enforcement appears to be the result of careless users, rather than a technical breach of the technology used to anonymize the currency transactions. For example, 10 people were arrested in the Netherlands in January 2016 as part of an international raid on online illegal drug markets, after they were caught converting bitcoins into euros in bank accounts using commercial bitcoin services, and then withdrawing millions in cash from ATM machines. Interpol and the U.S. Federal Bureau of Investigation were able to follow the trail of bitcoin addresses allegedly linking that money to online illegal drug sales, which were all recorded in the bitcoin blockchain.

## Not Ideal for B2B Crimes

The emphasis on cryptocurrencies may be misplaced, particularly with respect to identifying and tracking large criminal transactions, according to security experts. While small-time criminals and thrill-seekers often use bitcoin and oth-

er cryptocurrencies to transact on the dark web, experts say large money transactions have migrated to currencies that do not need to be exchanged (which open up the account holder to being identified) to be used in the real world.

“People focus on cryptocurrency, and focus on bitcoin,” says Scott Dueweke, president and founder of Zebryx Consulting, which focuses on anonymous transactions, digital forensics, and the Dark Web. However, he says the bulk of illicit money transactions are flowing through Russian-based electronic currency systems, such as WebMoney and Perfect Money.

“It’s very important to distinguish and [challenge] the notion that this is all about bitcoin, and that’s the primary driver of criminal activity,” Dueweke says. “It’s an important driver of the criminal underworld and buying illegal goods, and it’s well suited to the individual who is a casual purchaser.”

Dueweke notes big-time criminals usually choose to use types of Russian-backed currencies that are largely out of the reach of U.S. and other Western anti-money laundering and banking laws, rather than cryptocurrencies, which have their own issues (while one’s identity is obscured, there is a full record of all transactions on the blockchain, which could ultimately be used to trace back transactions if any one actor slips up and discloses his or her identity).

“If you’re really trying to make these purchases as part of the criminal marketplace, doing it through a system you know is immune to Western law enforcement, immune to the type of controls set up for the banking system, and is run, most likely, in some sort of collaboration with the [Russian] oligarchs and law enforcement to look the other way, that is a much better solution,” Dueweke explains.

Dueweke likens the choice in payment type to where each criminal lies on the food chain.

“If you’ve got the casual drug user, or small-time drug dealer trying to buy relatively small amounts to sell locally, yeah, he’s going to end up using bitcoin, and he may or may not use it effectively to avoid being traced,” Dueweke says. “But the guy on the back end, who is part of some drug cartel, if they have some sort of network for buying and selling at the B2B scale, that seems to be going on pre-

## Dueweke says the bulk of illicit money transactions flow through Russian-based electronic currency systems.

dominately using other digital payment types, or traditional movement mechanisms, such as trade-based money laundering, bulk cash, or stored value cards.”

### Funding Cyberattacks

Still, it is not just the purchase of illegal goods using cryptocurrency that has law enforcement and industry leaders worried. The availability of largely anonymous currency is also seen as helping to facilitate cybercrime and cyberattacks. The use of cryptocurrency as a payment type can also be exploited by those individuals and groups that conduct cyberattacks, as the sponsor of the attack can use cryptocurrency to pay those who carry out the attack, obscuring the money trail.

“There’s been a predominance of bitcoin use for ransomware campaigns,” says Ed Cabrera, chief security officer at Trend Micro. “[Criminals] want to make it as easy as possible to pay the ransom.”

In a ransomware attack, a company may be targeted with a denial-of-service attack or other breach, and then be required to make a payment in order to allow the company or user to regain access to their network or files. The use of cryptocurrencies as a payment mechanism, which obscures the recipient of the ransom payment, also may accelerate the use of so-called “zero-day” attacks, which exploit previously unknown technical vulnerabilities, thereby leaving security professionals with little or no time to prepare a patch or fix, leaving them no choice but to pay a ransom.

These types of attacks appear to be on the upswing. Trend Micro’s tracker on the number of ransomware attacks indicate 72 attacks were reported in the first half of 2016, up 172% from the previous year. In all of 2015, just

29 ransomware attacks were reported by Trend Micro.

Whether stopping illegal purchases on the Dark Web or trying to make it more difficult for bad actors to initiate and monetize Zero-Day attacks, experts believe the first step is to focus on better understanding the various types of cryptocurrencies used, their strengths and weaknesses, and where they are being exchanged into more liquid currencies.

“The first thing needs to be focusing on the cryptocurrencies,” Cabrera says. “Without changing any laws, I’d try to focus on these exchange houses. There are some that are criminally focused, and you can tell because they charge a high-end amount in administrative fees. They’re charging a higher fee; they’re pretty much providing protection.”

Thorough investigations and more stringent international money laundering laws may be a good first step in stopping some small-time purchasers and sellers on the Dark Web, but are unlikely to have an impact on those operating from geographic safe havens.

“There’s really nothing you can do about it, from a law enforcement perspective,” Dueweke says. “Typically, they’re set up in an area of the world where they’re economically and politically repressed, and they have relationships with local law enforcement for protection.”

Indeed, many of the bad actors simply thumb their noses at stringent international regulations and laws. Says Dueweke: “If you’re an exchanger in Pakistan, you’re laughing at the regulations.”

### Further Reading

Still Don’t Get Bitcoin? Here’s an Explanation Even a Five-Year-Old Will Understand, *CoinDesk*, January 9, 2014, <http://www.coindesk.com/bitcoin-explained-five-year-old/>

What Was Silk Road and How Did It Work?, *PC Magazine*, October 3, 2013, <http://www.pcmag.com/article2/0,2817,2425184,00.asp>

The Deep Web - Onion Routing, Tor, Dark Net Markets, Crypto Currencies Explained, <https://www.youtube.com/watch?v=5d1MGPQnWoU>

Keith Kirkpatrick is principal of 4K Research and Consulting, based in Lynbrook, N.Y.

© 2017 ACM 0001-0782/17/3 \$15.00



# ACM Recognizes New Fellows

**A**CM HAS RECOGNIZED 53 of its members as ACM Fellows for major contributions in areas including artificial intelligence, cryptography, computer architecture, high performance computing and programming languages. The achievements of the 2016 ACM Fellows are accelerating the digital revolution, and affect almost every aspect of how we live and work today.

“As nearly 100,000 computing professionals are members of our association, to be selected to join the top one percent is truly an honor,” explains ACM President Vicki L. Hanson. “Fellows are chosen by their peers and hail from leading universities, corporations and research labs throughout the world. Their inspiration, insights and dedication bring immeasurable benefits that improve lives and help drive the global economy.”

Underscoring ACM’s global reach, 2016 Fellows hail from organizations in Australia, Austria, Canada, China, France, India, Israel, Italy, The Netherlands, Switzerland, the United Kingdom and the United States.

The 2016 Fellows have been cited for numerous contributions in areas including cloud computing, computer security, data science, Internet routing and security, large-scale distributed computing, mobile computing, spoken-language processing and theoretical computer science.

ACM will formally recognize its 2016 Fellows at the annual Awards Banquet, to be held in San Francisco on June 24, 2017. Additional information about the 2016 ACM Fellows, the awards event, as well as previous ACM Fellows and award winners is available on the ACM Awards site at <http://awards.acm.org/>.

## 2016 ACM Fellows

**Noga Alon**, Tel Aviv University

**Paul Barford**, University of Wisconsin

**Luca Benini**, Swiss Federal Institute of Technology, Zurich and Università di Bologna

**Ricardo Bianchini**, Microsoft Research

**Stephen Blackburn**, Australian National University

**Dan Boneh**, Stanford University

**Carla E. Brodley**, Northeastern University

**Justine Cassell**, Carnegie Mellon University / Language Technologies Institute

**Erik Demaine**, Massachusetts Institute of Technology

**Allison Druin**, University of Maryland

**Fredo Durand**, Massachusetts Institute of Technology

**Nick Feamster**, Princeton University

**Jason Flinn**, University of Michigan

**William Freeman**, Massachusetts Institute of Technology

**Yolanda Gil**, University of Southern California

**Robert L. Grossman**, University of Chicago / Open Data Group

**Rajesh K. Gupta**, University of California, San Diego

**James Hendler**, Rensselaer Polytechnic Institute

**Monika Henzinger**, Universität Wien

**Tony Hey**, The Science and Technology Facilities Council’s Rutherford Appleton Laboratory

**Xuedong Huang**, Microsoft AI and Research

**Daniel Jackson**, Massachusetts Institute of Technology

**Robert J.K. Jacob**, Tufts University

**Somesh Jha**, University of Wisconsin

**Ravi Kannan**, Microsoft Research

**Anne-Marie Kermarrec**, Mediego/Inria

**Martin Kersten**, Centrum Wiskunde & Informatica

**Christoforos Kozyrakis**, Stanford University

**Marta Kwiatkowska**, University of Oxford

**James Landay**, Stanford University

**K. Rustan M. Leino**, Microsoft Research

**J. Bryan Lyles**, Oak Ridge National Laboratory

**Todd C. Mowry**, Carnegie Mellon University

**Trevor Mudge**, University of Michigan, Ann Arbor

**Sharon Oviatt**, Incaa Designs

**Venkata N. Padmanabhan**, Microsoft Research India

**Shwetak Patel**, University of Washington

**David Peleg**, The Weizmann Institute of Science

**Radia Perlman**, Dell-EMC

**Adrian Perrig**, ETH Zurich

**Ganesan Ramalingam**, Microsoft Research India

**Louisa Raschid**, University of Maryland

**Holly Rushmeier**, Yale University

**Michael Saks**, Rutgers, The State University of New Jersey

**Sachin S. Sapatnekar**, University of Minnesota

**Abigail Sellen**, Microsoft Research

**Sudipta Sengupta**, Microsoft Research

**André Seznec**, INRIA

**Valerie E. Taylor**, Texas A&M University

**Carlo Tomasi**, Duke University

**Paul Van Oorschot**, Carleton University

**Manuela M. Veloso**, Carnegie Mellon University

**Zhi-Hua Zhou**, Nanjing University

# ACM

## ON A MISSION TO SOLVE TOMORROW.

Dear Colleague,

Without computing professionals like you, the world might not know the modern operating system, digital cryptography, or smartphone technology to name an obvious few.

For over 60 years, ACM has helped computing professionals be their most creative, connect to peers, and see what's next, and inspired them to advance the profession and make a positive impact.

We believe in constantly redefining what computing can and should do.

ACM offers the resources, access and tools to invent the future. No one has a larger global network of professional peers. No one has more exclusive content. No one presents more forward-looking events. Or confers more prestigious awards. Or provides a more comprehensive learning center.

Here are just some of the ways ACM Membership will support your professional growth and keep you informed of emerging trends and technologies:

- Subscription to ACM's flagship publication *Communications of the ACM*
- Online books, courses, and videos through the **ACM Learning Center**
- Discounts on registration fees to ACM Special Interest Group conferences
- Subscription savings on specialty magazines and research journals
- The opportunity to subscribe to the **ACM Digital Library**, the world's largest and most respected computing resource

Joining ACM means you dare to be the best computing professional you can be. It means you believe in advancing the computing profession as a force for good. And it means joining your peers in your commitment to solving tomorrow's challenges.

Sincerely,



Vicki L. Hanson  
President  
Association for Computing Machinery



Association for  
Computing Machinery

*Advancing Computing as a Science & Profession*

# SHAPE THE FUTURE OF COMPUTING. JOIN ACM TODAY.

ACM is the world's largest computing society, offering benefits and resources that can advance your career and enrich your knowledge. We dare to be the best we can be, believing what we do is a force for good, and in joining together to shape the future of computing.

## SELECT ONE MEMBERSHIP OPTION

### ACM PROFESSIONAL MEMBERSHIP:

- Professional Membership: \$99 USD
- Professional Membership plus ACM Digital Library: \$198 USD (\$99 dues + \$99 DL)
- ACM Digital Library: \$99 USD (must be an ACM member)

### ACM STUDENT MEMBERSHIP:

- Student Membership: \$19 USD
- Student Membership plus ACM Digital Library: \$42 USD
- Student Membership plus Print *CACM* Magazine: \$42 USD
- Student Membership with ACM Digital Library plus Print *CACM* Magazine: \$62 USD

- Join ACM-W:** ACM-W supports, celebrates, and advocates internationally for the full engagement of women in computing. Membership in ACM-W is open to all ACM members and is free of charge.

Priority Code: CAPP

### Payment Information

Name \_\_\_\_\_  
ACM Member # \_\_\_\_\_  
Mailing Address \_\_\_\_\_  
City/State/Province \_\_\_\_\_  
ZIP/Postal Code/Country \_\_\_\_\_  
Email \_\_\_\_\_

Payment must accompany application. If paying by check or money order, make payable to ACM, Inc., in U.S. dollars or equivalent in foreign currency.

- AMEX    VISA/MasterCard    Check/money order

Total Amount Due \_\_\_\_\_  
Credit Card # \_\_\_\_\_  
Exp. Date \_\_\_\_\_  
Signature \_\_\_\_\_

### Purposes of ACM

ACM is dedicated to:

- 1) Advancing the art, science, engineering, and application of information technology
- 2) Fostering the open interchange of information to serve both professionals and the public
- 3) Promoting the highest professional and ethics standards

Return completed application to:  
ACM General Post Office  
P.O. Box 30777  
New York, NY 10087-0777

Prices include surface delivery charge. Expedited Air Service, which is a partial air freight delivery service, is available outside North America. Contact ACM for more information.

**Satisfaction Guaranteed!**

## BE CREATIVE. STAY CONNECTED. KEEP INVENTING.



Association for  
Computing Machinery

1-800-342-6626 (US & Canada)  
1-212-626-0500 (Global)

Hours: 8:30AM - 4:30PM (US EST)  
Fax: 212-944-1318

acmhelp@acm.org  
acm.org/join/CAPP



DOI:10.1145/3041043

Pamela Samuelson

# Legally Speaking

## Supreme Court on Design Patent Damages in *Samsung v. Apple*

*Considering influences leading to the recent U.S. Supreme Court decision in a years-long case that Apple filed against Samsung over iPhone design infringement.*

**S**HOULD SAMSUNG HAVE to pay Apple \$399 million—its total profits on sales of certain smartphones—for infringement of three Apple design patents when the value of the Samsung phones may be attributable to many other desirable features and not just to the designs infringed? An anomalous rule in U.S. design patent law seems to suggest the answer is yes, when it should be no.

The U.S. Supreme Court heard oral arguments to determine the answer to this question last October: the Court decided the case in early December 2016, when it threw out the nearly \$400 million in damages Samsung had been ordered to pay Apple by a ruling of the Court of Appeals for the Federal Circuit (CAFC) affirming an award of all of Samsung's profits from selling the infringing phones. The exact amount of damages, to be determined by the U.S. Court of Appeals for the Federal Circuit or a trial court on remand, will

likely be much less than the hundreds of millions of dollars Samsung might have paid.

Several technology companies (including Facebook, eBay, and Google) and technology industry associations filed amicus curiae (friend of the court) briefs. They warned that upholding the total profits award against Samsung would lead to a deluge of litigation and result in unjustified windfalls when design patents are infringed as to only one or a small number of components of complex multicomponent products. The Court's decision will have huge implications for the technology industry.

After providing some background on design patents, this column discusses the arguments that the litigants and the U.S. government took on the "total profits" issue and the way the Justices reacted to those arguments.

### Origins of the "Total Profits" Rule

Ornamental designs for articles of manufacture have been eligible for

design patent protection in the U.S. since 1842. Their inventors must apply to the Patent and Trademark Office (PTO), satisfy novelty and nonobviousness standards, and claim the design through drawings and descriptions of the article of manufacture to which the design will be applied. (Most other countries provide legal protection original designs of article of manufacture, although not through the patent system.) Design patent protections may now last for up to 15 years.

Design patents in the 19<sup>th</sup> and early 20<sup>th</sup> centuries conventionally covered simple articles of manufacture, such as carpets and wallpaper, which were attractive to consumers because of the patented design. Design patents today are more likely to be sought for designs applied to specific components of complex products.

Infringement of a design patent occurs when an unlicensed person embodies that design in an article of manufacture and the accused product is so

similar that an ordinary observer would be deceived into buying the infringer's product thinking it was buying the patentee's product. (Embodying the design in a different type of product generally does not infringe because consumers will not be deceived in this manner.)

When design patents have been infringed, courts may order defendants to pay the patentee a reasonable royalty for use of the patented design in infringing products. Alternatively, design patentees can ask for a disgorgement of the defendant's profits as to the article of manufacture to which that design has been applied. (Courts have ruled that design patentees cannot get both disgorgement of profits and reasonable royalties, as that would produce double recovery.)

### Origins of the "Total Profits" Rule

In the late 19<sup>th</sup> century, in two cases involving design patents for carpets, the U.S. Supreme Court gave a narrow interpretation to the disgorgement of profits rule. The Court denied the patentee an award of the infringers' profits because he had not proven how much of the infringers' profits were due to the patented design and how much was other factors (such as the quality of the wool).

In response to criticism of these decisions, the U.S. Congress in 1887 amended the design patent statute so that patentees could get the "total profits" that defendants derived from selling articles of manufacture embodying the patented designs. Congress was aware that this new "total profits" rule might overcompensate some patentees, but regarded this outcome as better than a rule that undercompensated them. There is, however, no comparable "total profits" rule in any other intellectual property law.

### Apple's Design Patents and Total Profits Award

Three design patents on the external configuration of smartphones were at issue in *Samsung*. One was for a black rectangular round-cornered front face for the device. A second was for a rectangular round-cornered front face with a surrounding rim or bezel. A third was for a colorful grid of 16 icons to be displayed on a screen.

In the trial Apple brought against Samsung for infringing these pat-



ents, the judge instructed the jury that it could not assess how much of Samsung's profits from selling smartphones were attributable to the patented designs. If the jury found infringement, it was obliged to award Samsung's total profits from sales of infringing phones. The jury agreed with Apple on the infringement claims and awarded \$399 million in total profits. Samsung appealed to the CAFC.

The appellate court acknowledged it was difficult to justify this award for infringement of the three Apple design patents as a matter of equity. However, the CAFC decided the statute required it to affirm the total profits award because it regarded Samsung's smartphones as the relevant "article

of manufacture" to which the patented designs had been applied.

Samsung persuaded the Supreme Court to review the CAFC ruling.

### Solicitor General Weighs In

The U.S. government rarely files briefs with Supreme Court cases or joins in oral argument when disputes are between private litigants such as Apple and Samsung. The Solicitor General (SG) filed a brief in *Samsung* to challenge the CAFC's ruling, saying it would result in "grossly excessive and essentially arbitrary awards" for design patent infringement in cases in which the patented design was applied to one component of a multicomponent product (such as a latch for a refrigerator-

tor door). The SG also participated in the oral argument to represent the government's interest in sound interpretations of U.S. design patent law.

The SG argued that the proper inquiry in cases involving multicomponent products was, first, to identify the relevant "article of manufacture" to which the patented design(s) had been applied, and second, to assess what portion of the defendant's profits were attributable to the infringing article. In respect of multicomponent products, the relevant article of manufacture may be one component, rather than the product as a whole, even though there may be no separate market for that component.

The SG's brief identified several factors that juries should take into account in deciding what the relevant article of manufacture was: the scope of the patented design; the extent to which the patented design was responsible for the appeal of the product; the existence of other conceptually distinct and unrelated components of the product; and how various components of the product were manufactured.

The SG recognized that it would sometimes be difficult for the jury to determine what "total profits" were attributable to the infringing components, but regarded the design patent statute as requiring this determination. The SG also recognized that when components embodying patented designs were not sold separately, the total profits inquiry would be "functionally similar" to the conventional profits-attributable-to-infringement analysis used in other types of IP cases. However, the SG stated that "a significant conceptual and practical difference [exists] between the profit attributable to the infringing *article* and the profit attributable to the *infringement*" (emphasis in the original).

Profits attributable to the infringing article will generally be higher than profits attributable to infringement, especially when the relevant article of manufacture is valuable for more than the design. (Samsung's lawyer suggested, for instance, that the design-patented rectangular round-edge design for smartphones might be valuable to consumers because it makes the face less likely to fracture in addition to making the phone look "cool.") Total

## The Justices did not discuss Apple's patents at all or the allocation of profits to the smartphones at issue.

profits on the round-edge component may overcompensate Apple, but this must be what Congress intended when it amended the law in 1887.

The SG recommended sending the Samsung case back to the lower courts to determine the relevant article of manufacture and profits attributable to that article under this standard.

### The Supreme Court Argument

Before the Court, nobody defended the CAFC's ruling that juries must award total profits on the sale of products embodying patented designs. Samsung pointed to hundreds of thousands of component parts in smartphones and argued that the three patented designs were only small components of the smartphones at issue. In a new trial on damages, Samsung argued that the jury should first study the patent to examine the design and the article to which the design was applied. The jury should then make a judgment about the profits attributable to the components embodying the infringing designs. Consumer surveys and expert witnesses might help the jury to decide these issues.

Although not defending the CAFC ruling, Apple asked the Supreme Court to affirm the total profits award against Samsung. It argued that its patented designs made its smartphones "peculiar and distinctive in appearance," as patented designs often do. (Judges sometimes decide that lower courts erred in their interpretation of a legal rule, but find the error to be too insubstantial to justify a new trial.)

Most of the Justices' questions focused on the difficulties that juries would have in deciding what the relevant article of manufacture was and

how much of the profits from the overall product should be attributable to the component in which the infringing design had been embodied.

Not all cases would be difficult, however. Total profits on products such as wallpaper and carpets should be easy insofar as the patented designs drove consumer demand for the product. Also relatively easy would be cases in which the patents were for small components of complex multicomponent products (for example, designs for car cup holders, windshields for boats, or hood ornaments for cars).

The Justices did not discuss Apple's patents at all or the allocation of profits to the smartphones at issue. But they speculated about what juries might do in allocating profits for infringement of a hypothetical design patent covering the overall shape of a car such as the Volkswagen Beetle. Some Justices seemed to think that consumer demand for cars embodying this design would be near the total profits for the car as a whole, while other Justices thought that much of the value of such a car would lie in the mechanical and other functional design elements not covered by that design patent.

### Conclusion

Although the Supreme Court oral argument in *Samsung* largely focused on non-technology design patent examples, the Justices were very aware of the concerns raised by many technology companies and industry associations about the deleterious effects of excessive awards in design patent litigation posed by the CAFC's total-profits-on-products ruling.

The Court provided very little guidance in its *Samsung* decision about how fact-finders should assess the relevant article of manufacture to which patented designs have been applied and the profits attributable to that article in its *Samsung* ruling. We can all breathe a sigh of relief that the worst outcome of the case has been averted by the Court's willingness to correct yet another erroneous ruling by the CAFC. □

**Pamela Samuelson** (pam@law.berkeley.edu) is the Richard M. Sherman Distinguished Professor of Law and Information at the University of California, Berkeley, and a member of the ACM Council.

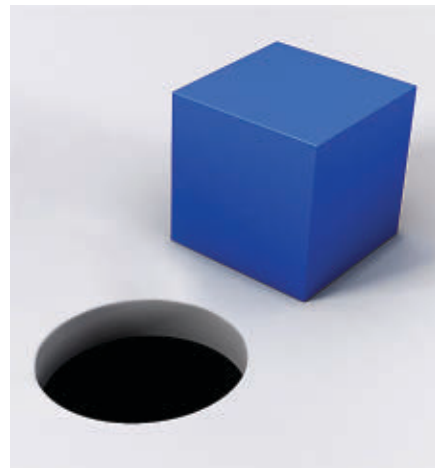
Copyright held by author.

## Computing Ethics Where Review Goes Wrong

*Examining professional misconduct among academic publication examiners.*

**I** AM A researcher twice accused of professional misconduct in the publication process. The first incident happened when I was a junior professor submitting the definitive paper from my Ph.D. research to a journal. The second happened quite recently. Despite these accusations, I am a successful researcher and teacher. This column is my appeal to reviewers and editors for caution and moderation.

In the first instance I went through several rounds of reviews, revisions, and resubmissions. All but one of the reviewers accepted the paper, and the paper was eventually rejected. I resubmitted the paper to another journal. Unbeknownst to me, the reviewer who had previously rejected the paper was contacted as reviewer again. The result was that the editor, in an email sent to all reviewers, charged me with knowingly submitting a paper with incorrect results. It had never occurred to me that I was doing anything wrong. I felt scared, helpless, ashamed, alone, and confused. It took a while to dig out proof that I had checked the veracity of the paper. I forwarded to the editor all previous reviews and my responses. I also forwarded my email correspondences with a mathematics researcher who had helped me verify proofs and address the reviewer's concerns. The review process was restarted with the same set of four reviewers; as I expected, the paper was rejected. I rarely submitted to a journal again because I was



terrified of being charged with trying to “shop” a rejected paper.

The second accusation of professional misconduct happened recently. Some months ago, my student and I submitted a paper to a conference. A couple of months later, we submitted new research—generated since the last submission—to a workshop connected to the conference. The conference paper was rejected; instead of getting reviews for the workshop paper, we were accused of unethical conduct for submitting a paper that had significant overlap with another paper in review, and for not citing the paper in review. No proof or examples of the overlap were presented. There was no attempt to contact us authors prior to the accusation and negative decision concerning our workshop submission. We tried to contest the decision by sending proof that the writing overlap between

the two papers was 3% and by explaining that the shorter workshop paper presented a new algorithm developed and validated after submission of the earlier paper. The conference chairs and the program committee served as the judge and the jury; we had no arbitrators, no voice. We were punished without being given a chance to rebut.

Trying to publish a paper that was rejected earlier or trying to publish new results are part of the publication process, and researchers should not fear being charged with professional misconduct for doing their job. The review process is unbalanced with concentration of power in the hands of reviewers. In both incidents, when a reviewer charged wrongdoing, there was an immediate presumption of guilt followed by punishment. By the time we were contacted, our guilt and punishment was fait accompli. We felt helpless and wronged, with no possibility of our names being cleared of wrongdoing. When a reviewer suspects something amiss, it is important that editors contact the authors for an explanation. The final decision should reflect input from both sides. In order for authors to understand and accept a decision, they should feel that their voices were heard.

The review process should incorporate the ethos of research and the publication process. A review process is adversarial since reviewers are tasked with checking that a paper is correct, relevant, and original, while

authors believe that their paper is correct, relevant, and original. Both reviewers and authors may make mistakes, but errors/misjudgments by reviewers can lead to punishment of authors. Therefore, it is important for editors/chairs to be impartial, which is not possible without author input.

A reviewer spends a few hours on a paper, while authors invest several years, so authors often understand their research more than the reviewer. When a paper is rejected at one venue, authors resubmit to another venue, hoping that a new set of reviewers accept their paper. Authors try to revise their paper based on reviewer comments, but it is sometimes impossible to address all the reviewers' concerns. Possible reasons for not addressing comments are: lack of time since the next conference deadline follows immediately, author fatigue after going through countless revisions, lack of resources to address comments, contradictory comments by various reviewers, to name a few. Therefore, reviewers who are reassigned papers they rejected elsewhere should not label authors as unethical if their comment is unaddressed in this new submission.

Sometimes rules are ambiguous and authors unintentionally break a rule. Apropos, rules on when and how to cite one's previous papers are contradictory: for double-blind, citing one's older papers is wrong; for single-blind, not citing one's older papers is wrong. When a paper is resubmitted, it is possible that authors forget to add/remove their paper citation. Sometimes, authors are simply embarrassed by their earlier paper and choose not to cite it. Sometimes, authors do not cite older papers since it appears as an attempt to increase citation count of their papers. Reviewers may attribute sinister intentions where none exists. Instead of charging authors with misconduct, asking authors for an explanation is reasonable. Editors/chairs may subsequently choose to eliminate the paper from consideration without charging authors with professional misconduct.

Reviewing is subject to error and bias.<sup>1,2</sup> In both incidents, the reviewing was single-blind. The double-blind review process is vanishing, which is

## The review process is unbalanced with concentration of power in the hands of reviewers.

unfortunate. A reviewer might be biased by author names and affiliations, so removal of the double-blind process hurts researchers from lower-ranked institutions. Moreover, some conferences allow reviewers to resubmit their review after seeing other submitted reviews of the paper; this exacerbates the problem of biased review. Authors who are not part of the elite group have to scale an impossibly high bar to get their research published in reputed conferences/journals.

In both incidents, charges were brought by reviewers who had rejected an earlier submission by the authors. Before punishing us, the possibility of reviewer fatigue and bias should have been considered. A friend who is on several program committees laughingly mentioned that he rejected a paper, submitted to three different conferences, thrice. The authors of this paper may have given up without realizing the paper was reviewed by the same set of reviewers. This problem could be addressed by asking potential reviewers whether they have previously reviewed (and rejected) the paper, and if so, whether they could impartially review this new submission. The framing of these questions might help reviewers understand their biases. If impartiality is not possible, then one should recuse oneself from reviewing the paper.

Reviewing research papers is a difficult task, and I thank reviewers, editors, program chairs, and others involved in the process. It is arduous to read and comprehend technical papers that are likely written by young researchers who are learning to articulate their research. In recent years, paper submissions have spiked, so reviewers may be reading a large number of papers.

Time pressure on reviewers has led to non-reviews, which provide little to no feedback. Sometimes, reviews are sprinkled with words such as "moronic," "stupid," or "myopic" that reflect the frustration of reviewers. One solution might be to increase the size of program committees by making them more inclusive and to reduce overlap in program committees. Reviewer bias could be reduced by bringing back double-blind reviews and by ensuring that, as a general rule, reviewers do not review papers they have earlier rejected at another venue.

Almost a decade after the first incident, the editor who accused me of misconduct sought me out and apologized. I thank this editor because his apology allowed me to evaluate and acknowledge the impact of that first wrongful accusation. This second accusation has little impact on my career. I am speaking up on behalf of young researchers who are just embarking on their careers. I appeal for rules and guidelines, which protect David and keep Goliath in check. I appeal for the psychology of research to be incorporated into the review process, for program committees to have more diversity and less overlap, for reviewers to understand their inherent biases, and above all, for chairs and editors to have a preponderance of evidence before charging authors with wrongdoing. After all, if Hardy had accused the unknown young mathematician who sent him well-known theorems as original work, one of the greatest mathematical geniuses, Ramanujan, would have been lost to the world. **□**

### References

1. Publication Ethics by Council of Scientific Editors, 2016; <http://www.councilscienceeditors.org/resource-library/editorial-policies/white-paper-on-publication-ethics/>
2. Kahneman, D. *Thinking, Fast, and Slow*. Farrar, Straus and Giroux, 2011.

**Elizabeth Varki** (varki@cs.unh.edu) is an associate professor in the Department of Computer Science at the University of New Hampshire.

I thank Rachelle Hollander for her guidance through the review process and for steering this column toward publication. I thank several anonymous reviewers for their comments and suggestions, which have greatly improved the column; special thanks to the reviewer who suggested the column title. Finally, I thank Cindy Childers, Larry Dowdy, Andras Fekete, Anu Mathai, and Royce Withey for their support and feedback on early versions of this column.

Copyright held by author.



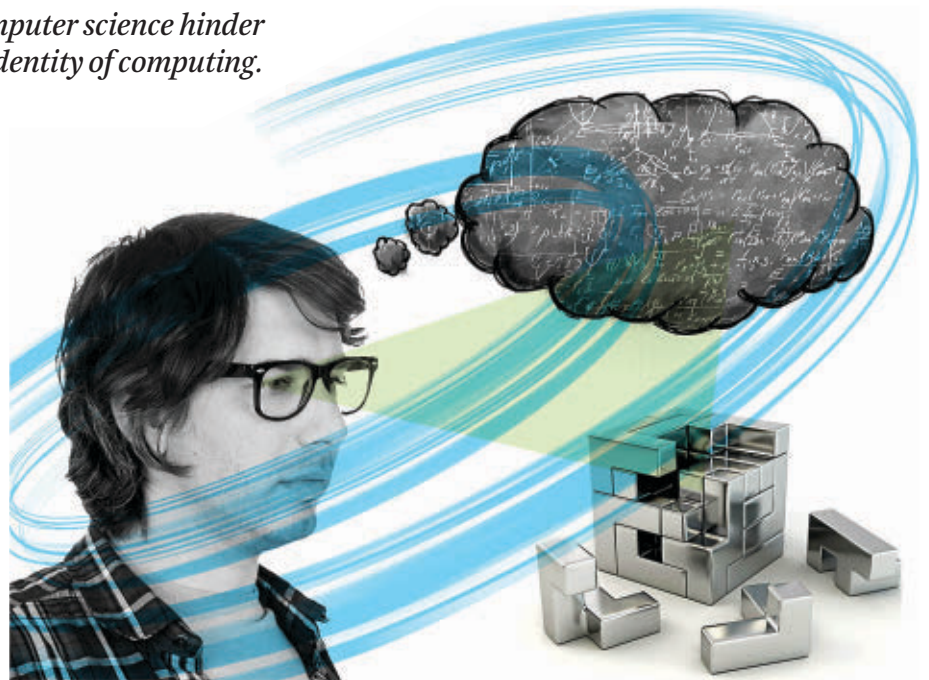
## The Profession of IT Misconceptions About Computer Science

*Common misconceptions about computer science hinder professional growth and harm the identity of computing.*

**W**HEN MANY OF US were in school, we were given definitions of computer science such as “the study of information processes and their transformations” or “the study of phenomena arising around computers.” But when we entered the world of professional practice, we experienced computer science in a completely different way from these abstract definitions.

In our professional world, our ability to obtain a job depends on how well we display competence in using computational methods and tools to solve problems of interest to our employers. We have to be able to create small apps on the fly with no more effort than writing a Post-It note. We discover that we have customers who can be satisfied or not with our work—and that our professional advancement depends on an ever-expanding legacy of satisfied customers. We discover that over time we become proficient and our peers and bosses call on us to solve ever more complex problems. We are beset with unpredictable surprises and contingencies not covered in school or our previous experience—and yet we must deal with them effectively.

As an example, the current surge of deep-learning AI technologies has generated many benefits and created well-paying new jobs for data analysts and software designers who automate some mental tasks. These technologies are permanently displacing workers who used to do those tasks manually.



Many readers of this column are well-paid designers and yet even they worry that a technology surprise might push them overnight into the unemployed. Our Internet technology has facilitated globalization of labor and raised living standards everywhere, yet has stimulated a backlash of anti-immigration, anti-trade sentiment. Our Internet technology has also developed a dark side that includes hackers, data and identity thieves, scammers, polarizing websites, terrorists, and more. To help us cope with all this change and churn we have organized ourselves into several hundred professional specialty groups hosted by ACM, IEEE, and others.

Because computing is so intimately involved with many fields, an educational movement called “CS for All”

has emerged that aims to include some computing in everyone’s K–12 education or professional development.

We note that the CS for All movement does not advocate that every single child should learn to program for the sake of becoming a professional programmer or software engineer. Computing occupations are projected to grow at a higher rate than all other STEM areas combined. By one estimate, more than 7.7 million Americans use computers in complex ways in their jobs, almost half of them in fields that are not directly related to STEM.<sup>a</sup> Regardless of their career, many professionals will be using computer science at work.

<sup>a</sup> <https://advancesinap.collegeboard.org/stem/computer-science-principles>

We have worked closely with many people in this movement. They have been confronted with a number of misconceptions about computer science, both in the audiences they are trying to reach and among themselves. These misconceptions can lead to expectations that cannot be met—for example, graduates thinking they have studied things that will help them land and keep good jobs, or employer expectations about what graduating professionals can do for them. These misconceptions can also interfere with practitioner abilities to navigate effectively in the real world of computing. Our purpose here is to call out the nine most pernicious of these misconceptions and call on our professional colleagues to work to dispel them.

**CS = programming.** The idea that programming is the core activity of computer science is easy to accept and yet it is only partly true. Computing professionals are expected to be able to program. But computing professionals engage in many other important activities such as designing software and hardware systems, networks, databases, and applications. The idea that coding (a subset of programming) opens the door to many career opportunities has intrigued the public because of the successful publicity of Hour of Code, after-school coding clubs for boys and girls, and coding competitions.

This misconception is not new. It took root in the 1970s and was repeatedly challenged over the years; ACM and IEEE, for example, spent considerable effort in the 1990s uprooting it.<sup>7</sup> The most recent ACM/IEEE college curriculum includes 17 areas of computing technology besides programming.<sup>2</sup> Even when computing is distilled to its core scientific and engineering principles it is still a huge field in which programming is not the lead.<sup>4</sup> The new Advanced Placement course CS Principles<sup>b</sup> reflects a much broader view of computer science for high school seniors. Code.org's K-12 curriculum<sup>c</sup> covers much more than coding. Yet the “learn to code” movement seems to offer quick access to many well-paying jobs after you work your way through intensive bootcamps and workshops. The moment of truth comes when you dis-

cover in interviews that employers look for much more than ability to code.

**Programming is concerned with expressing a solution to a problem as notation in a language.** The purpose of programs is to control machines—not to provide a means of algorithmic self-expression for programmers. Starting with Ada Lovelace's example programs in the 1840s, programming has always been concerned with giving instructions to a machine so that the machine will produce an intended effect. A programming language is a notation used to encode an algorithm that, when compiled into executable code, instructs a machine.

Computer scientists have long understood that every programming language (the “syntax”) is bound to an abstract machine (the “semantics”). The machine—simulated by the compiler and operating system on the real hardware and network—carries out the work specified by the algorithm encoded in the program. Advanced programmers go further: they design new abstract machines that are more suited to the types of problems needing solution. The idea that programs are simply a notation for expression is completely disconnected from this fundamental reality that programs control machines.

A recent illustration of this is the legal battle by copyright owners to block the distribution of decryption software that unlocked copyright protection. The decryption software would have been uninteresting if it were merely a means of expression. But that software, when run on a machine, broke the copy protection.

**Programmers progress from beginners to experts over a long period of time. The much publicized kid coders are mostly beginners.**

**Once you master a core knowledge base including variables, sequencing, conditionals, loops, abstraction, modularization, and decomposition, you will be a computing professional.** This is a woefully incomplete characterization of what computing professionals need to know. The concepts listed are all programming concepts, and programming is a small subset of CS. The listed concepts *were* central in the 1960s and 1970s when programming was the main interface to computers. Today, you simply *cannot* be a competent programmer with little skill at systems, architectures, and design, and with little knowledge of the domain in which your software will be used.

**Programming is easy to learn.** Programming and coding are skill sets. Programmers can progress from beginners to experts over a long period of time. It takes more and more practice and experience to reach the higher stages. Becoming proficient at programming real-world applications is not easy. The much publicized kid coders are mostly beginners.

Educators have been searching for many years for ways to accelerate learning programming. Seymour Papert introduced the Logo language in the 1970s and watched how children got excited by computing and learned how to think computationally. He attuned Logo to children's interests; even so, it still took students time to move from the fascination of the introduction to the ability to program useful computations regularly.

**Computational thinking is the driver of programming skill.** Computational thinking (CT) is an old idea in CS, first discussed by pioneers such as Alan Perlis in the late 1950s.<sup>8</sup> Perlis thought “algorithmizing” would become part of every field as computing moved in to automate processes. Dijkstra recognized he had learned new mental skills while programming (1974). In his 1980 book *Mindstorms*, Papert was the first to mention the term CT explicitly when discussing the mental skills children developed while programming in Logo. Jeannette Wing catalyzed a discussion about how people outside CS could benefit from learning computing.<sup>8</sup> The common thread was always that CT is the *consequence* of learning to program.

Modern versions of the CT story have turned this upside down, claim-

b <https://code.org/educate>

c [http://www.csteachers.org/?page=CSTA\\_Standards](http://www.csteachers.org/?page=CSTA_Standards)

ing that CT is a knowledge set that drives the programming skill. A student who scores well on tests to explain and illustrate abstraction and decomposition can still be an incompetent or insensitive algorithm designer. The only way to learn the skill is to practice for many hours until you master it. The newest CSTA guidelines move to counteract this upside-down story, emphasizing exhibition of programming skill in contests and projects.<sup>d</sup>

Because computation has invaded so many fields, and because people who do computational design in those fields have made many new discoveries, some have hypothesized that CT is the most fundamental kind of thinking, trumping all the others such as systems thinking, design thinking, logical thinking, scientific thinking, etc. This is computational chauvinism. There is no basis to claim that CT is more fundamental than other kinds of thinking.

**When we engage in everyday step-by-step procedures we are thinking computationally.** Everyday step-by-step procedures use the term “step” loosely to refer to an isolated action of a person. That meaning of step is quite different from a machine instruction; thus most “human executable recipes” cannot be implemented by a machine. This misconception actually leads people to misunderstand algorithms and therefore overestimate what a machine can do.

Step-by-step procedures in life, such as recipes, do not satisfy the definition of algorithm because not all their steps are machine executable. Just because humans can simulate some computational steps does not change the requirement for a machine to do the steps. This misconception undermines the definition of algorithm and teaches people the wrong things about computing.

**Computational thinking improves problem-solving skills in other fields.** This old claim is called the “transfer hypothesis.” It assumes that a thinking skill automatically transfers into other domains simply by being present in the brain. It would revolutionize education if true. Education researchers have studied automatic transfer of CT for three decades and have never been able

to substantiate it.<sup>6</sup> There is evidence on the other side—slavish faith in a single way of thinking can make you into a worse problem solver than if you are open to multiple ways of thinking.

Another form of transfer—designed transfer—holds more promise. Teachers in a non-CS field, such as biology, can bring computational thinking into their field by showing how programming is useful and relevant in that field. In other words, studying computer science alone will not make you a better biologist. You need to learn biology to accomplish that.

**CS is basically science and math. The engineering needed to produce the technology is all based on the science and math.** History tells us otherwise. Electrical engineers designed and built the first electronic computers without knowing any computer science—CS did not exist at the time. Their goal was to harness the movement of electrons in circuits to perform logical operations and numerical calculations. Programs controlled the circuits by opening and closing gates. Later scientists and mathematicians brought rigorous formal and experimental methods to computing. To find out what works and what does not, engineers tinker and scientists test hypotheses. In much of computing the engineering has preceded the science. However, both engineers and scientists contribute to a vibrant computing profession: they need each other.

**Old CS is obsolete. The important developments in CS such as AI and big data analysis are all recent.** Computing technology is unique among technologies in that it sustains exponential growth (Moore’s Law) at the levels of individual chips, systems, and economies.<sup>3</sup> Thus it can seem that computer technology continually fosters upheavals in society, economies, and politics—and it obsoletes itself every decade or so. Many of the familiar principles of CS were identified in the 1950s and 1960s and continue to be relevant today. The early CS shaped the world we find ourselves in today. Our history shows us what worked and what does not. The resurrection of the current belief that CS=programming illustrates how those who forget history can repeat it.

Artificial intelligence is an old subfield of CS, started in the early 1950s. For

the first 30 years, AI pursued a dream of intelligent machines. When they were unable to even get close to realizing the dream, they gave up rule-based AI systems and turned instead to machine learning focused on automating simple mental tasks rather than general intelligence. They were able to build amazing automations based on neural networks without trying to imitate human brain processes. Today’s AI has become so successful with neural network models that do far better than humans at some mental tasks that we are now facing social disruptions about joblessness caused by AI-driven automation.

## Conclusion

We welcome the enthusiasm for computer science and its ways of thinking. As professionals, we need to be careful that in our enthusiasm we do not entertain and propagate misconceptions about our field. Let us not let others oversell our field. Let us foster expectations we can fulfill. ■

## References

1. Change the Equation. The hidden half. Blog post. (Dec. 7, 2015); <http://changetheequation.org/blog/hidden-half>
2. Computer Science Curricula 2013; <https://www.acm.org/education/CS2013-final-report.pdf>
3. Denning, P. and Lewis, T.G. Exponential laws of computing growth. *Commun. ACM* 60, 1 (Jan. 2017).
4. Denning, P. and C. Martell. *Great Principles of Computing*. MIT Press, 2015.
5. Denning, P.J. et al. Computing as a discipline. *Commun. ACM* 32, 1 (Jan. 1989), 9–23.
6. Guzdial, M. *Learner-Centered Design of Computing Education: Research on Computing for Everyone*. Morgan-Claypool, 2015.
7. Tedre, M. *Science of Computing: Shaping a Discipline*. CRC Press, Taylor & Francis, 2014.
8. Tedre, M. and Denning, P.J. The long quest for computational thinking. In *Proceedings of the 16<sup>th</sup> Koli Calling Conference on Computing Education Research* (Koli, Finland, Nov. 24–27, 2016), 120–129.
9. Wing, J. Computational thinking. *Commun. ACM* 49, 3 (Mar. 2006), 33–35.

**Peter J. Denning** (pjd@nps.edu) is Distinguished Professor of Computer Science and Director of the Cebrowski Institute for information innovation at the Naval Postgraduate School in Monterey, CA, is Editor of *ACM Ubiquity*, and is a past president of ACM. The author’s views expressed here are not necessarily those of his employer or the U.S. federal government.

**Matti Tedre** (matti.tedre@acm.org) is Associate Professor of Computer and Systems Sciences at Stockholm University, Sweden, adjunct professor at University of Eastern Finland, and the author of *Science of Computing: Shaping a Discipline* (CRC Press, Taylor & Francis, 2014).

**Pat Yongpradit** (pat@code.org) is the Chief Academic Officer for Code.org and served as staff lead on the development of the K–12 Computer Science Framework. A former high school computer science teacher, Pat has been featured in the book, *American Teacher: Heroes in the Classroom*, has been recognized as a Microsoft Worldwide Innovative Educator, and is certified in biology, physics, math, health, and technology education.

Copyright held by authors.

<sup>d</sup> One of the K–12 curriculum recommendations actually cites making a peanut butter and jelly sandwich as an example of an algorithm.

## Viewpoint

# Learning with Mobile Technologies

*Considering the challenges, commitments, and quandaries.*

**W**ITH A FEW years of hindsight, the previously ambitious but now notorious rollout of iPads by the Los Angeles Unified School District certainly looks “spectacularly foolish.”<sup>5</sup> Quite consistently, researchers, industry experts, journalists, school personnel, and politicians agree the plan was well intentioned, but ill conceived and doomed from the start. They lament that if the school district had a more comprehensive blueprint for selecting, using, and managing the technology, the enterprise would have been successful. This cycle of hype and disappointment continues to characterize large-scale adoptions of technology in schools across the globe.<sup>2,12</sup> The accompanying lessons, however, are surprisingly short-lived. I recently attended an international forum with participants from across groups of stakeholders and the message was quite clear: technology in schools equals innovation; let’s not waste time being negative about technology; let’s just get on with it. Such a cavalier approach to learning technologies in schools and the flippant reaction to any cautions and critiques only serve to further jeopardize the learning opportunities of students who have been historically marginalized in schools.

This Viewpoint presents my reflections on struggles encountered in a curricular reform project that relied heavily on new technologies in the classroom.<sup>7-11</sup> I am transparent about



the difficulties we experienced in the hope that our candor will allow for pause and deliberation as others embark on similar efforts, ultimately providing them a more advantageous point of departure. Recognizing the importance of place and context, I do not expect that our challenges will be identical to what others face across varied learning environments. That said, I sincerely hope that strong proponents of the “just get on with it” position will have the courage to not dismiss our concerns as idiosyncratic.

Our project was deployed in high school classrooms throughout the Los

Angeles area over the last six years. A defining characteristic of the reform effort, implemented in high school computer science, data science, math, and science classrooms, was to have students use mobile technologies to collect data about themselves and about issues that were important to them. The collection and the analysis of personally relevant data were intended to promote computational and statistical thinking in STEM.

### Challenges

From a learning technology perspective, we faced three top-level chal-

lenges. First, a considerable amount of instructional time was squandered dealing with technology issues. Particularly in large classrooms with a single teacher, precious days of instructional time were lost trying to ensure the technology worked with different platforms, with devices that ran older software, and with network specifications that varied across sites. Similar to the phenomenon identified by Hasu and Engeström,<sup>4</sup> these troubles emerged because the idealized conditions in which technology is developed rarely match the messy conditions in which it is actually used. Not only do technology developers struggle to anticipate real-world challenges, they fail to recognize them and empathize with users when bugs, errors, and user troubles begin to manifest.

Second, the novelty effect of mobile phones not only waned but gradually morphed into a source of student opposition.<sup>7,8</sup> The project assumed that mobile phones would motivate students to collect and analyze their own data. The use of phones soon got repetitive and lost its allure. In fact, many students eventually resented having to complete assignments that required smartphones. Without adequate attention to the pedagogy that would sustain interest and learning, mobile technologies became a hindrance to student engagement.

Third, the mobile app and the corresponding desktop-based software were often not responsive to students' developing interests. This issue was perhaps more pronounced since our technology was meant to engage students as producers rather than just consumers of data—an emphasis that required tailored software. But, students started asking questions they could not adequately answer with our platform. Given the large investment of time and resources, there was at least an implicit pressure to continue to use the technology. Within a dynamic that allowed technology to supersede teachers' and students' creativity and inventiveness, the technology eventually started to constrain student learning.

### Commitments

Our biggest lesson is that the success of any classroom learning technology requires a deep commitment to valu-

## Profound dilemmas about the use of mobile technology emerged the deeper we engaged with issues of implementation.

ing the expertise, creativity, goals, and desires of teachers and students. Such a commitment is particularly demanding and laborious since it calls on us to design technology and learning experiences *with* rather than *for* teachers and students as they live and learn in the richness and complexities of their contexts.

The first commitment must be to students. Technologies must be used to create learning opportunities that build on students' strengths and dynamic interests and recognize their emergent hopes and goals.<sup>3</sup> But, adult assumptions about youth can make such learner-centered approaches to technology difficult. As I have documented, adults assume time and again that young people's out-of-school interests will transfer fluidly into school-based learning.<sup>7,8</sup> We have shown that utilizing technologies on the presumption they are a part of youth culture can backfire. Young people can come to resent that their out-of-school interests are co-opted and appropriated in the curriculum. For instance, in many of the classrooms we observed, students began to feel burdened by the smartphones they used for school, going so far as to say they were no longer "phones" but "devices for school." They were more than happy to Instagram a picture of a particularly delicious meal, but were aggravated and exasperated that teachers required them to document their snacks with smartphones for the data analysis component of the curriculum. In a particularly striking case, students were indignant that financial resources were expended

# Calendar of Events

### March 4–5

I3D '17: Symposium on Interactive 3D Graphics and Games, San Francisco, CA, Sponsored: ACM/SIG, Contact: Kenny Mitchell, Email: drkennymitchell@yahoo.com

### March 6–9

HRI '17: ACM/IEEE International Conference on Human-Robot Interaction, Vienna, Austria, Contact: Bilge Dincer Mutlu, Email: bilge@cs.wisc.edu

### March 7–11

CHIIR '17: Conference on Human Information Interaction and Retrieval, Oslo, Norway, Sponsored: ACM/SIG, Contact: Ragnar Nordlie, Email: ragnar.nordlie@hioa.no

### March 13–16

IUI'17: 22<sup>nd</sup> International Conference on Intelligent User Interfaces, Limassol, Cyprus, Co-Sponsored: ACM/SIG, Contact: George Angelos Papadopoulos, Email: george@cs.ucy.ac.cy

### March 16–17

TAU '17: ACM International Workshop on Timing Issues in the Specification and Synthesis of Digital Systems, Bay Area, CA, Sponsored: ACM/SIG, Contact: Qiuyang Wu, Email: qiuyang.wu@gmail.com

### March 19–22

ISPD '17: International Symposium on Physical Design, Portland, OR, Sponsored: ACM/SIG, Contact: Mustafa Ozdal, Email: mustafa.ozdal@cs.bilkent.edu.tr

### March 20–23

TEI '17: 10<sup>th</sup> International Conference on Tangible, Embedded, and Embodied Interaction, Yokohama, Japan, Sponsored: ACM/SIG, Email: inakage@kmd.keio.ac.jp

on mobile technologies while what mattered to them most, like band and music, were cut at their schools. Students were exceedingly frustrated that adults and outsiders made superficial assumptions about what would engage them rather than valuing students' real, context-specific desires and aspirations.

The second commitment must be to teachers. We must unsettle problematic discourses that attempt to explain the failure of technological innovation through teacher resistance or complacency. We need to shift our focus away from "training" teachers to use particular technologies. Rather, we need to start from a place that respects teachers' professional expertise and, from there, facilitate the space for teachers to access and leverage technologies as one set of tools in a repertoire of pedagogical resources. As I have argued elsewhere,<sup>7</sup> technologies should be considered in light of the texts, tools, and talk they make available for teaching and learning. Technology in the classroom is successful when pedagogy is effective. We must learn to trust and support teachers as they closely consider the possibilities and limitations of technologies in their specific contexts and decide to leverage them (or not leverage them) accordingly.

### Quandaries

Profound dilemmas about the use of mobile technology emerged the deeper we engaged with issues of implementation.<sup>8</sup> First, who will provide the technology? If schools provide them, how do we address issues of liability? How can policies be formulated so they do not limit access to students if they or their parents are unable or unwilling to assume liability? If the assumption is that students use their own devices, what expectations arise for students to purchase and possess up-to-date, compatible devices? Similar challenges emerge in terms of data plans.

Second, a recurring issue that came up in our work was that students felt they lacked freedom with mobile technologies in schools. Schools are often required to limit access to websites, social networking sites, and modes of digital communication. But these limits change the very meaning of


digital technologies for students and inadvertently discourage their usage for school-sanctioned, instructional purposes. The proposition to simply let students use mobile technologies is naive. Mobile technologies cannot be an island of freedom in an otherwise controlling and constricting learning environment. Such contradictions lead to mobile technologies as a source of disruption and subversion by students. We must do the hard work to make schools places where students are trusted with their own learning. It is within a larger commitment to respecting students' agency that mobile technologies can lead to expanded learning opportunities.

Third, as students use mobile technologies, they generate large amounts of data, consciously and more often without explicit consideration or awareness. Such data can be powerful when used to customize learning experiences for students. But, the same data can inadvertently limit possible trajectories of learning through dynamics analogous to the "filter bubble."<sup>6</sup> Additionally, given the realities of funding and liability, one could easily imagine arrangements where corporations provide technologies to schools in exchange for access to students' data—deals that are already in place. These trends raise weighty and far-reaching questions about the purpose of schooling in a democratic society.

### Conclusion

If classrooms, schools, and society are inequitable, the introduction of mobile technologies into classroom spaces will not fundamentally alter these inequities. Equitable learning does not simply transpire through disruptive innovation that uses technology. We need to engage in the difficult work of understanding and addressing relationships of power, authority, and knowledge in the classroom. We must create learning environments where students and their cultural practices are valued and built upon. We need spaces where students feel connected to their peers and adults. We must nurture classrooms where students engage in democratic deliberation about issues of equity and justice.<sup>9</sup> We need to address the soci-

etal inequities and injustices in which schooling is embedded. If we simultaneously address these needs, mobile technologies can benefit all students. Otherwise, as history has shown us, the introduction of new technologies in classrooms will continue, for the most part, to reproduce existing patterns of success and failure.<sup>1</sup>

Mobile technologies have permeated our society. There is no question whether we should incorporate them into schools or not. They are already in schools and will most likely become a more significant part of our daily lives, both in and out of schools. To those who say, "let's just get on with it," I say certainly. But let's do it in a manner that deeply wrestles with the challenges and quandaries of mobile technologies, and in ways that honor the complexities of teaching and learning and respects the agency of teachers and students. Else, the "just do it" attitude will get us nowhere. 

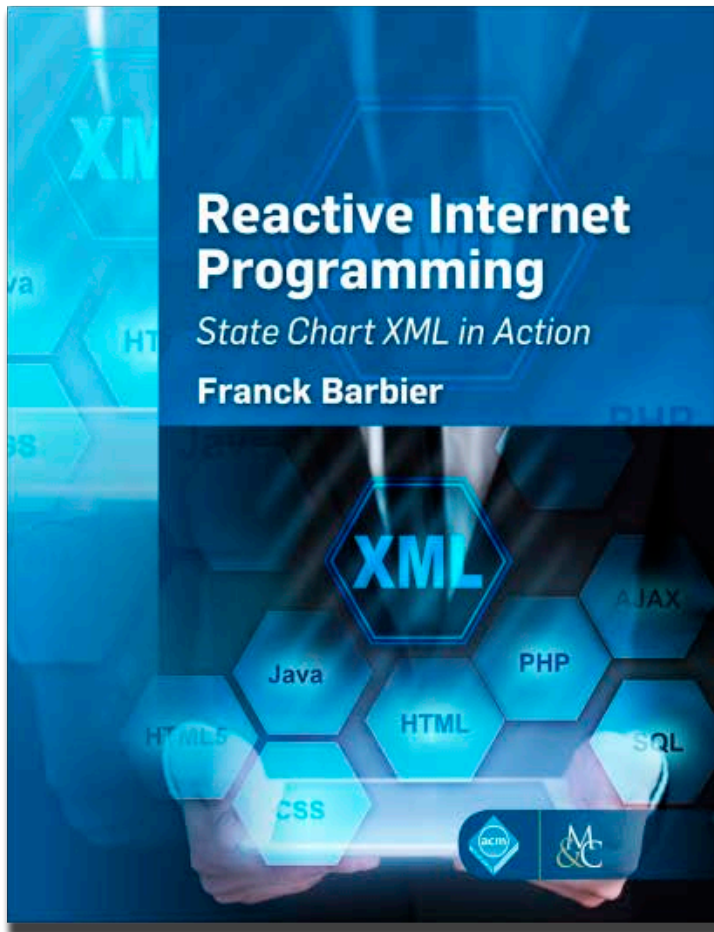
### References

1. Cuban, L. *Oversold and Underused: Computers in the Classroom*. Harvard University Press, Cambridge, MA, 2001.
2. Graham, M. Kenya's laptops for schools dream fails to address reality. *The Guardian*. (June 27, 2013).
3. Guzdial, M. *Learner-Centered Design of Computing Education: Research on Computing for Everyone*. Morgan & Claypool Publishers, San Rafael, CA, 2016.
4. Hasu, M. and Engeström, Y. Measurement in action: An activity-theoretical perspective on producer/user interaction. *International Journal of Human-Computer Studies* 53, 1 (Jan. 2000), 61–89.
5. Lapowsky, I. What schools must learn from LA's iPad debacle. *Wired* (May 8, 2015).
6. Pariser, E. *The Filter Bubble: What the Internet is Hiding from You*. Penguin Press, New York, 2011.
7. Philip, T.M. and Garcia, A. The importance of still teaching the iGeneration: New technologies and the centrality of pedagogy. *Harvard Educational Review* 83, 2 (2013), 300–319.
8. Philip, T.M. and Garcia, A. Schooling mobile phones: Assumptions about proximal benefits, the challenges of shifting meanings, and the politics of teaching. *Educational Policy* 29, 4 (2015), 676–707.
9. Philip, T.M., Olivares-Pasillas, M.C., and Rocha, J. Becoming racially literate about data and data literate about race: A case of data visualizations in the classroom as a site of racial-ideological micro-contestations. *Cognition & Instruction* 34, 4 (2016), 361–388.
10. Philip, T.M., Schuler-Brown, S., and Way, W. A framework for learning about Big Data with mobile technologies for democratic participation: Possibilities, limitations, and unanticipated obstacles. *Technology, Knowledge and Learning* 18, 3 (2013), 103–120.
11. Philip, T.M. et al. When educators attempt to make "community" a part of classroom. *Teacher Education* 34 (2013), 174–183.
12. Tashobya, A. Government to revamp one laptop per child programme. *The New Times*. (Mar. 19, 2015).

Thomas M. Philip (tmp@ucla.edu) is an associate professor in the Graduate School of Education and Information Studies at UCLA.

This work was partially supported by a National Science Foundation grant (MSP-0962919).

Copyright held by author.



You need effective means to put reactive programming into practice.

**THEY'RE  
RIGHT HERE**

Is Internet software so different from “ordinary” software? This book practically answers this question through the presentation of a software design method based on the State Chart XML W3C standard along with Java. Web enterprise, Internet-of-Things, and Android applications, in particular, are seamlessly specified and implemented from “executable models.”

Internet software puts forward the idea of event-driven or reactive programming, as pointed out in Bonér et al.’s “Reactive Manifesto”. It tells us that reactivity is a must. However, beyond concepts, software engineers require effective means with which to put reactive programming into practice. *Reactive Internet Programming* outlines and explains such means.

The lack of professional examples in the literature that illustrate how reactive software should be shaped can be quite frustrating. Therefore, this book helps to fill in that gap by providing in-depth professional case studies that contain comprehensive details and meaningful alternatives. Furthermore, these case studies can be downloaded for further investigation.

Internet software requires higher adaptation, at run time in particular. After reading *Reactive Internet Programming*, you will be ready to enter the forthcoming Internet era.



<http://books.acm.org>

<http://www.morganclaypoolpublishers.com/xml>

Article development led by [acmqueue](https://queue.acm.org)  
queue.acm.org

## A computing adventure about time through the looking glass.

BY THEO SCHLOSSNAGLE

# Time, but Faster

EVERY ONCE IN a while, you find yourself in a rabbit hole, unsure of where you are or what time it might be. This article presents a computing adventure about time through the looking glass.

The first premise was summed up perfectly by the late Douglas Adams in *The Hitchhiker's Guide to the Galaxy*: "Time is an illusion. Lunchtime doubly so." The concept of time, when colliding with decoupled networks of computers that run at billions of operations per second, is ... well, the truth of the matter is you simply never really know what time it is. That is why Leslie Lamport's seminal paper on Lamport timestamps was so important to the industry, but this article is actually about wall-clock time, or a reasonably useful estimation of it.

Even on today's computers, it is feasible to execute an instruction in less than a nanosecond. When the white rabbit looks at his pocket watch in *Alice's Adventures in Wonderland*, he is seeing what time it was a nanosecond before, as the light travels from the hands on the watch to his eye—assuming that Lewis Carroll's timeless tale took place in a vacuum and that the rabbit was holding the watch one-third of a meter from his eyes.

When you think of a distributed system where a cluster of fast computers are often more than one light-nanosecond away from each other, it is understandably difficult to time something that starts in one place and ends in another with nanosecond precision; this is the realm of physicists, not bums like us with commodity computing hardware run in environments we don't even manage. To upset the newcomer even further, every computer today is effectively a distributed system itself, with each CPU core having its own clock ticking away, with its own subtly different frequency and sense of universal beginning.

All that said, computers must give users the illusion of a clock. Without it, we won't know what time it is. As computers get faster and faster, we are able to improve the performance of our systems, but one fundamental of performance engineering is that we cannot improve what we cannot measure; so measure we must. The fundamental paradox is that as what we measure gets smaller, the cost of measuring it remains fixed, and thus becomes relatively monstrous.

### The Beginning of the Tumble ...

At Circonus, we write a database that is fast and keeps getting faster. We dump energy into this seemingly endless journey because we operate at scale and every bit of efficiency we eke out results in lower COGS (cost of goods sold) for us and better service for our customers. Moreover, it fundamentally affords a cost effectiveness of telemetry collection and analysis that approaches reasonable



economics to “monitor all the things.” In that context ...

Let’s assume we want to achieve an average latency for operations of one microsecond. Now let’s wrap some numbers around that. I will make some notes about certain aspects of hardware, but I’ll really focus only on hardware from the past several years. While we like to think in terms of seconds, computers don’t care about this concept of time. They care only about clock ticks.

### The TSC

Online CPUs are forever marching forward at some frequency, and the period of this frequency is called a tick. In an effort to save power, many computers can shift between different power-saving states that cause the frequency of the CPU to change. This could make it excruciatingly difficult to tell high-granularity time accurately, if the frequency of the CPU were used for timing. Each core on a modern CPU has a TSC (time-stamp counter) that counts the number of ticks that have transpired. You can read this counter with the cleverly named `rdtsc` assembly instruction. Also, modern CPUs have a feature called an invariant TSC, which guarantees that the frequency at which ticks occur will not change for any reason (but mainly for power-saving mode changes). My development box has an invariant TSC that ticks approximately 2.5999503 times per nanosecond. Other machines have different frequencies.

The standard tooling to figure out how long an operation takes on a Unix machine is either `clock_gettime(CLOCK_MONOTONIC,...)` or `gethrtime()`. These calls return the number of nanoseconds since some arbitrary fixed point in the past. The examples shown here use `gethrtime()` because it is shorter to write.

```
hrtime_t start = gethrtime();
some_operation_worthy_of_measurement();
hrtime_t elapsed = gethrtime() - start;
```

As these things are measured, the



`gethrtime()` call itself will take some time. The question is: where does the time it returns sit relative to the beginning and end of the `gethrtime()` call itself? That can be answered with benchmarks. The bias introduced by measuring the start and finish is relative to its contribution to overall running time. In other words, if the “operation” being measured is made to take a long time over many iterations, the measurement bias can be reduced to near zero. Timing `gethrtime()` with `gethr-`

`time()` would look like this:

```
#define LOTS 1000000
hrtime_t start = gethrtime();
for(int i=0;i<LOTS;i++) (void)gethrtime();
hrtime_t elapsed = gethrtime() - start;
double avg_ns_per_op = (double) elapsed / (double)LOTS;
```

Behold, a benchmark is born. Furthermore, you could actually measure the number of ticks elapsed in the

test by bracketing the test with calls to `rdtsc` in assembly. Note that you must bind yourself to a specific CPU on the box to make this effective because the TSC clocks on different cores do not have the same concept of “beginning.” Table 1 shows the results if this is run on our two primary platforms (Linux and Illumos/OmniOS on a 24-core 2.6GHz Intel box).

The first observation is that Linux optimizes both of these calls significantly more than OmniOS does. This has actually been addressed as part of the LX brand work in SmartOS by Joyent and will soon be upstreamed into Illumos for general consumption by OmniOS. Alas, that isn’t the worst thing: objectively determining what time it is, is simply too slow for microsecond-level timing, even at the lower 119.8ns/op (nanoseconds per operation) number above. Note that `gettimeofday()`

supports only microsecond-level accuracy and thus is not suitable for timing faster operations.

At just 119.8 ns/op, bracketing a one-microsecond call will result in:

$$(119.8*2)/(1000 + 119.8*2) \rightarrow 19.33\%$$

So 19.33% of the execution time is spent on calculating the timing, and that doesn’t even include the time spent recording the result. A good goal to target here is 10% or less. So, how do we get there?

### Looking At Our Tools

These same modern CPUs with invariant TSCs have the `rdtsc` instruction, which reads the TSC, yet doesn’t provide insight into which CPU you are executing on. That would require either prefixing the call with a `cpuid` instruction or binding the executing

thread to a specific core. The former adds ticks to the work; the latter is wholly inconvenient and can really defeat any advanced NUMA (nonuniform memory access)-aware scheduling that the kernel might provide. Basically, binding the CPU provides a super-fast but overly restrictive solution. We just want the `gethrtime()` call to work and be fast.

We are not the only ones in need. Out of the generally recognized need, the `rdtscp` instruction was introduced. It supplies the value in the TSC and a programmable 32-bit value. The operating system can program this value to be the ID of the CPU, and a sufficient amount of information is emitted in a single instruction. Don’t be deceived; this instruction isn’t cheap and measures in at 34 ticks on this machine. If you code that instruction call as `uint64_t mtev_rdtscp(int *cpuid)`, that returns the TSC and optionally sets a `cpuid` to the programmed value.

The first challenge here is to understand the frequency. This is a straightforward timing exercise illustrated in the accompanying figure.

This usually takes around 10ns, assuming no major page fault during the assignment—10ns to set a piece of memory! Remember, that includes the average time of a call to `mtev_rdtscp()`, which is just over 9ns. That’s not really the problem. The problem is that sometimes we get HUGE answers. Why? We switch CPUs and the outputs of the two TSC calls are reporting two completely unrelated counters. So, to rephrase the problem: we must relate the counters.

The code for skew assessment is a bit much to include here. The basic idea is that we should run a calibration loop on each CPU that measures `TSC*nanos_per_tick` and assess the skew from `gethrtime()`, accommodating the running time of `gethrtime()`. As with most calibration loops, the most skewed is discarded and the remaining is averaged. This basically goes back to secondary-school math to find the linear intercept equation:  $y = mx + b$ , or:

$$\text{gethrtime}() = \text{nanos\_per\_tick} * \text{mtev\_rdtscp}() + \text{skew}$$

As the TSC is per CPU, you need to track  $m$  and  $b$  (`nanos_per_tick` and `skew`) on a per-CPU basis.

**Table 1. Starting benchmarks.**

Operating System	Call	Call Time
Linux 3.10.0	<code>gettimeofday</code>	35.7 ns/op
Linux 3.10.0	<code>gethrtime</code>	119.8 ns/op
OmniOS r151014	<code>gettimeofday</code>	304.6 ns/op
OmniOS r151014	<code>gethrtime</code>	297.0 ns/op

**Table 2. Results**

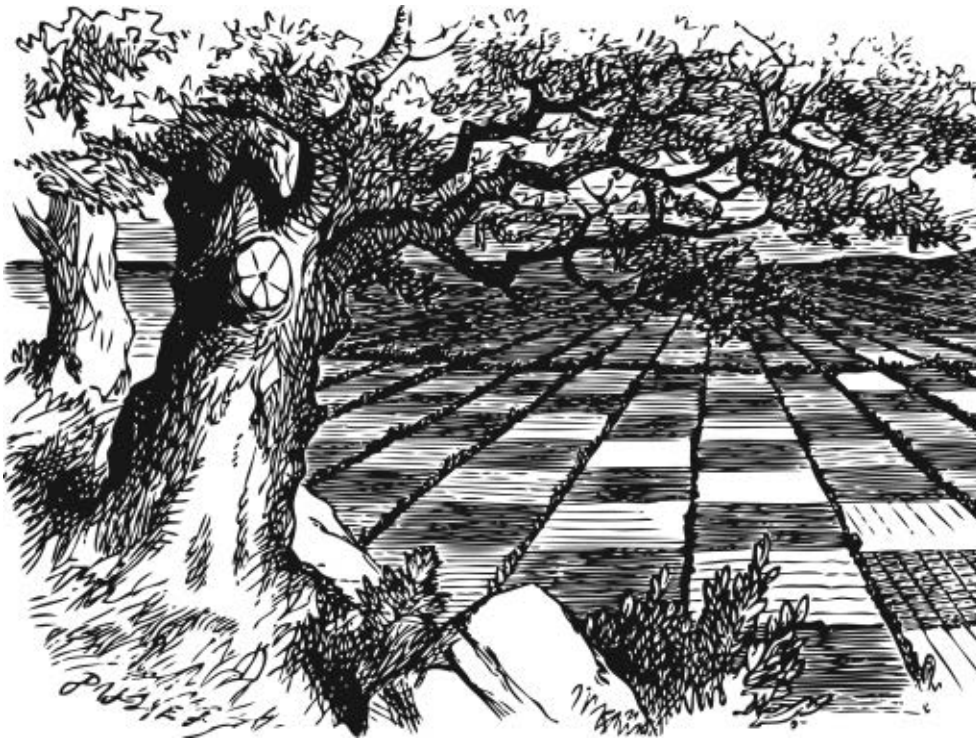
Operating System	Call	System Call Time	Mtev-variant Call	Speedup
Linux 3.10.0	<code>gettimeofday</code>	35.7 ns/op	35.7 ns/op	x1.00
Linux 3.10.0	<code>gethrtime</code>	119.8 ns/op	40.4 ns/op	x2.96
OmniOS r151014	<code>gettimeofday</code>	304.6 ns/op	47.6 ns/op	x6.40
OmniOS r151014	<code>gethrtime</code>	297.0 ns/op	39.9 ns/op	x7.44

### Calibration loop.

```
mtev_thread_bind_to_cpu(0);
hrtime_t start_ns = gethrtime();
uint64_t start_ticks = mtev_rdtscp(NULL);
sleep(10);
hrtime_t end_ns = gethrtime();
uint64_t end_ticks = mtev_rdtscp(NULL);
double nanos_per_tick = (double)(end_ns - start_ns) / (double)(end_ticks - start_ticks);
```

The next challenge becomes quite clear when testing this solution for timing the execution of a job—even the simplest of jobs:

```
uint64_t start = mtev_rdtscp(NULL);
*some_memory = 0;
uint64_t elapsed = mtev_rdtscp(NULL) - start;
```



Another nuance is that these two values together describe the translation between a CPU's TSC and the system's `gethrtime()`, and they are estimates. That means two important things: They need to be updated regularly to correct for error in the calibration and estimation; and they need to be set and read atomically. This is where the `cmpxchg16b` instruction enters.

Additionally, this calibration work is performed every five seconds in a separate thread, and we attempt to make that thread high priority on a real-time scheduler. It turns out that this all works quite well, even without the ability to change priority or scheduling class.

### Gravy

Since we're clearly having to correct for skew to align with the system `gethrtime()`, and the point in the past to which `gethrtime()` is relative is arbitrary (according to the documentation), we've elected to make that "arbitrary" point the Unix epoch. No additional instructions are required, and now the replacement `gethrtime()` can be used to power `gettimeofday()`. Therefore,  $y = mx + b$  is actually implemented as:

```
nano_second_since_epoch =
```

```
nanos_per_tick * mtev_rdtscp() + skew
```

Obviously, we'll pick up changes to the wall clock (via `adjtime()` et al.) only when we recalibrate.

### Safety

Obviously, things can and do go wrong. A variety of fail-safe mechanisms are in place to ensure proper behavior when the optimizations become unsafe. By default, if the lack of an invariant TSC is detected, the system is disabled. If a calibration loop fails for too long (15 seconds), the CPU is marked as bad and disabled. During rudimentary performance tests, if the system's `gethrtime()` can beat the emulation, then we disable. If all those tests pass, we still check to see if the underlying system can perform `gettimeofday()` faster than we can emulate it; if so, we disable `gettimeofday()` emulation. The goal is for `mtev_gethrtime()` to be as fast as or faster than `gethrtime()` and for `mtev_gettimeofday()` to be as fast as or faster than `gettimeofday()`.

### Results

The overall results are better than expected. The original goal was simply to provide a way for our implementation on Illumos to meet the performance


of Linux. The value of ZFS is deeply profound, and while Linux has some performance advantages in specific arenas, that doesn't matter much if you have undetectable corruption of the data you store.

Further optimization is possible in the implementation, but we've stopped for now, having achieved the initial goals. Additionally, for the purposes of this test, we have built the code portably. We can find a couple of nanoseconds if we compile `-march=native` on machines supporting the AVX (Advanced Vector Extensions) instruction set.

It is true that an approximately 40ns `gethrtime()` can be considered slow enough, relative to microsecond-level efforts, that very prudent selection is still necessary. It is also true that 40ns `gethrtime()` can open up a new world of possibilities for user-space instrumentation. It has certainly opened our eyes to some startling things.

### Acknowledgment

This all comes for free with [https://github.com/circonus-labs/libmtev/blob/master/src/utils/mtev\\_time.com](https://github.com/circonus-labs/libmtev/blob/master/src/utils/mtev_time.com) (see [https://github.com/circonus-labs/libmtev/blob/master/src/utils/mtev\\_time.c](https://github.com/circonus-labs/libmtev/blob/master/src/utils/mtev_time.c) for reference).

As of this writing, Linux and Illumos are supported platforms, and Darwin and FreeBSD do not have "faster time" support. The faster time support in `libmtev` was a collaborative effort between Riley Berton and Theo Schlossnagle. 

### Related articles on [queue.acm.org](http://queue.acm.org)

**Passively Measuring TCP Round-Trip Times**  
*Stephen D. Strowes*  
<http://queue.acm.org/detail.cfm?id=2539132>

**The One-Second War (What Time Will You Die?)**  
*Poul-Henning Kamp*  
<http://queue.acm.org/detail.cfm?id=1967009>

**Principles of Robust Timing over the Internet**  
*Julien Ridoux and Darryl Veitch*  
<http://queue.acm.org/detail.cfm?id=1773943>

**Theo Schlossnagle** is the founder and chief executive officer at Circonus, where he works on large-scale numerical data analysis. He is the author of *Scalable Internet Architectures* (Sams Publishing, 2006) and founder of OmniIT, an Internet consultancy.

Copyright held by owner/author.  
Publication rights licensed to ACM. \$15.00.

Article development led by [acmqueue](http://queue.acm.org)  
queue.acm.org

## Hardware and software perspectives.

BY MOHAMED ZAHRAN

# Heterogeneous Computing: Here to Stay

MENTIONS OF THE phrase *heterogeneous computing* have been on the rise in the past few years and will continue to be heard for years to come, because heterogeneous computing is here to stay. What is heterogeneous computing, and why is it becoming the norm? How do we deal with it, from both the software side and the hardware side? This article provides answers to some of these questions and presents different points of view on others.

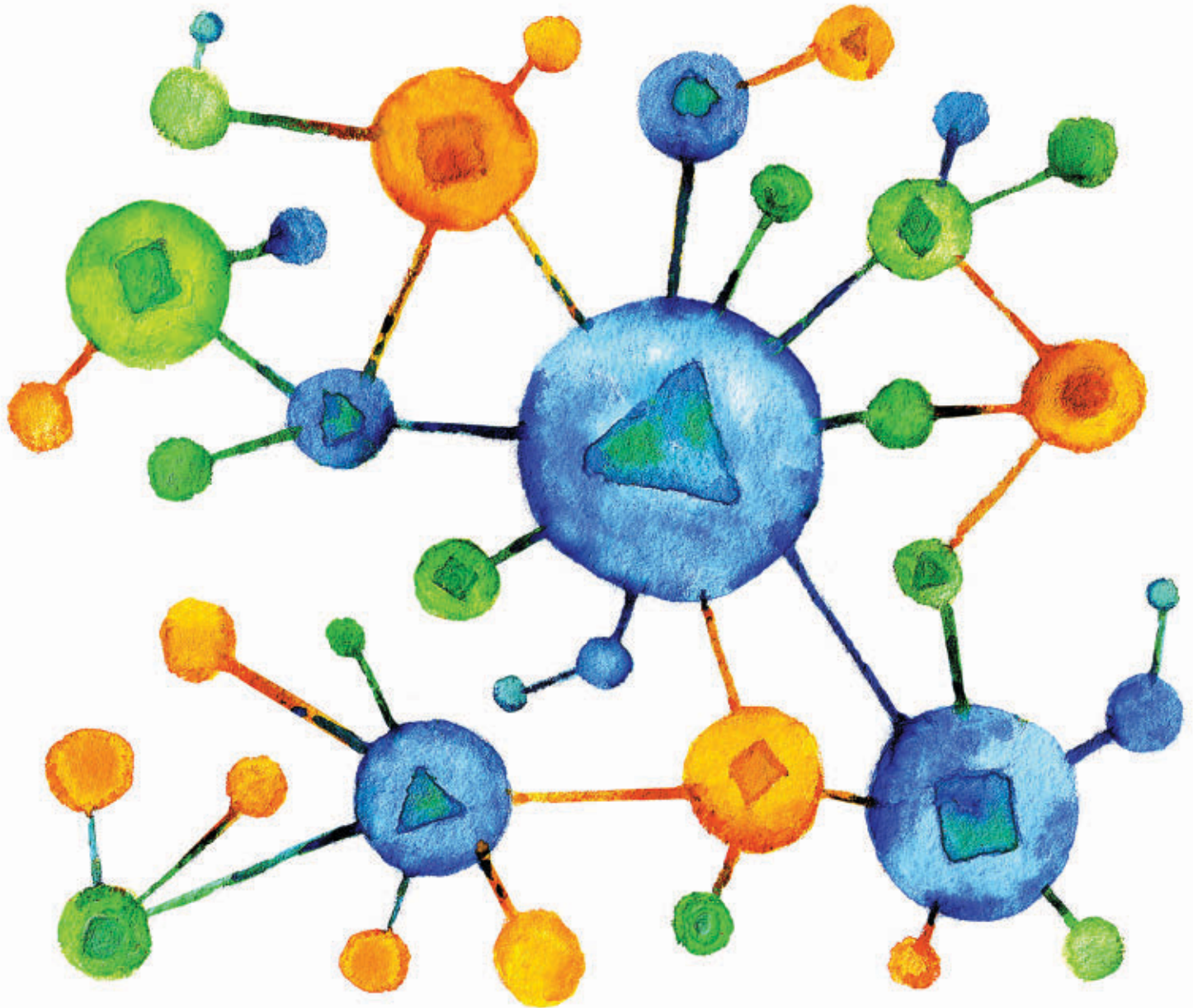
Let's start with the easy questions. What is heterogeneous computing? In a nutshell, it is a scheme in which the different computing nodes have different capabilities and/or different ways of executing instructions. A heterogeneous system is therefore a parallel system (single-core systems are almost ancient history). When multicore systems appeared, they were homogeneous—that is, all cores were similar. Moving from sequential programming to parallel programming, which used to be an area only for niche programmers, was a big jump. In heterogeneous computing, the cores are different.

Cores can have the same architectural capabilities—for example, the same hyperthreading capacity (or lack thereof), same superscalar width, vector arithmetic, and so on. Even cores that are similar in those capabilities, however, have some kind of heterogeneity. This is because each core now has its own DVFS (dynamic voltage and frequency scaling). A core that is doing more work will be warmer and hence will reduce its frequency and become, well, slower. Therefore, even cores with the same specifications can be heterogeneous. This is the first type of heterogeneity.

The second type involves cores with different architectural capabilities. One example is a processor with several simple cores (for example, single-issue, no out-of-order execution, no speculative execution), together with a few *fat* cores for example, with hyperthreading technology, wide superscalar cores with out-of-order and speculative execution).

These first two types of heterogeneity involve cores with the same execution model of sequential programming—that is, each core *appears* to execute instructions in sequence even if under the hood there is some kind of parallelism among instructions. With this multicore machine, you may write parallel code, but each thread (or process) is executed by the core in a seemingly sequential manner. What if computing nodes are included that don't work like that? This is the third type of heterogeneity.

In this type of heterogeneity the computing nodes have different execution models. Several different types of nodes exist here. The most famous is the GPU (graphics processing unit), now used in many different applications beside graphics. For example, GPUs are used a lot in deep learning, especially the training part. They are also used in many scientific applications and are delivering performance that is orders of magnitude better than traditional cores. The reason for this performance boost is that a GPU uses the single-instruction (or thread), mul-



multiple-data execution model. Let's assume you have a large matrix and need to multiply each element of this matrix by a constant. With a traditional core, this is done one element at a time or, at most, a few elements at a time. With a GPU, you can multiply all the elements at once, or in a very few iterations if the matrix is very large. The GPU excels in similar independent operations on large amounts of data.

Another computing paradigm that deviates from the traditional sequential scheme is the FPGA (field-programmable gate array). We all know that software and hardware are logically equivalent, meaning what you can do with software you can also do with hardware. Hardware solutions are much faster but inflexible. The FPGA

tries to close this gap. It is a circuit that can be configured by the programmer to implement a certain function. Suppose you need to calculate a polynomial function on a group of elements. A single polynomial function is compiled to tens of assembly instructions. A FPGA is a good choice if the number of elements needed to calculate the function is not large enough to require a GPU, and not small enough to be done in a traditional core efficiently. FPGAs have been used in many high-performance clusters. With Intel's acquisition last year of Altera, one of the big players in the FPGA market, tighter integration of FPGAs and traditional cores is expected. Also, Microsoft has started using FPGAs in its datacenter (Project Catapult).

A new member recently added to the computing-node options is the AP (Automata processor) from Micron.<sup>3</sup> AP is very well suited for graph analysis, pattern matching, data analytics, and statistics. Think of it as a hardware regular expressions accelerator that works in parallel. If you can formulate the problem at hand as a regular expression, then you can expect to get much higher performance than a GPU could provide. AP is built using FPGAs but designed to be more efficient in regular expressions processing.

Aside from the aforementioned computing nodes, there are many other processing nodes such as the DSP (digital signal processor) and ASIC (application-specific integrated circuit). Those target small niches of applica-

tions, however, and are not as versatile as the ones mentioned earlier. Brain-inspired neuromorphic chips, such as IBM's TrueNorth chip, are starting an era of cognitive computing.<sup>2</sup> Cognitive computing, championed by IBM's Watson and TrueNorth, is now used, after the impressive performance of the AI computer system Watson on "Jeopardy," in medical applications, and other areas are being explored. It is a bit early, however, to compare it with the other more general-purpose cores.

The rest of this article considers only traditional cores (with different capabilities), GPU, FPGA, and AP. The accompanying figure shows the big picture of a heterogeneous computing system, even though, because of the cost of programmability, finding a system with the level of heterogeneity shown in the figure is unlikely. A real system will have only a subset of these types.

What is the advantage of having this variety of computing nodes? The answer lies in performance and energy efficiency. Suppose you have a program with many small threads. The best choice in this case is a group of small cores. If you have very few complicated threads (for example, complicated control-flow graphs with pointer-chasing), then sophisticated cores (for example, fat superscalar cores) are the way to go. If you assign the complicated threads to simple cores, the result is poor performance. If you assign the simple threads to the sophisticated cores, you consume more power than needed. GPUs have

very good performance-power efficiency for applications with data parallelism. What is needed is a general-purpose machine that can execute different flavors of programs with high performance-power efficiency. The only way to do this is to have a heterogeneous machine.<sup>3</sup> Most machines now, from laptops to tablets to smart phones, have heterogeneous architectures (several cores and a GPU), and more heterogeneity is expected in the (very) near future. How should we deal with this paradigm shift from homogeneity to heterogeneity?

### Hardware Challenges

Several challenges exist at the hardware level. The first is memory hierarchy. The memory system is one of the main performance bottlenecks in any computer system. While processors had been following Moore's Law until a few years ago, making good leaps in performance, memory systems have not. Thus, there is a large performance gap between processor speed and memory speed. This problem has existed since the single-core era. What makes it more challenging in this case is the shared memory hierarchy (several levels of cache memory followed by the main memory). Who shares each level of caches? Each of the computational cores discussed here targets a program (or thread or process) with different characteristics from those targeted by other computational cores. For example, a GPU requires higher bandwidth, while a

traditional core requires faster access. As a result, what is needed is a memory hierarchy that reduces interference among the different cores, yet deals efficiently with the different requirements of each.

Designing such a hierarchy is far from easy, especially considering that, beside performance issues, the memory system is a nontrivial source of power consumption. This challenge is the subject of intensive research in industry and academia. Moreover, we are coming close to the era of nonvolatile memory. How can it best be used? Note here the heterogeneity in memory modules: for caches (SRAM), volatile memory (DRAM), nonvolatile memory (MRAM, STT-RAM, PCM, ReRAM, and many more technologies).

Another challenge at the hardware level is the interconnect: How should we connect the different cores and memory hierarchy modules? Thick wires dissipate less power but result in lower bandwidth because they take more on-chip space. There is a growing body of research in optical interconnect. The topology (ring, torus, mesh), material (copper, optical), and control (network-on-chip protocols) are hot topics of research at the chip level, at the board level, and across boards.

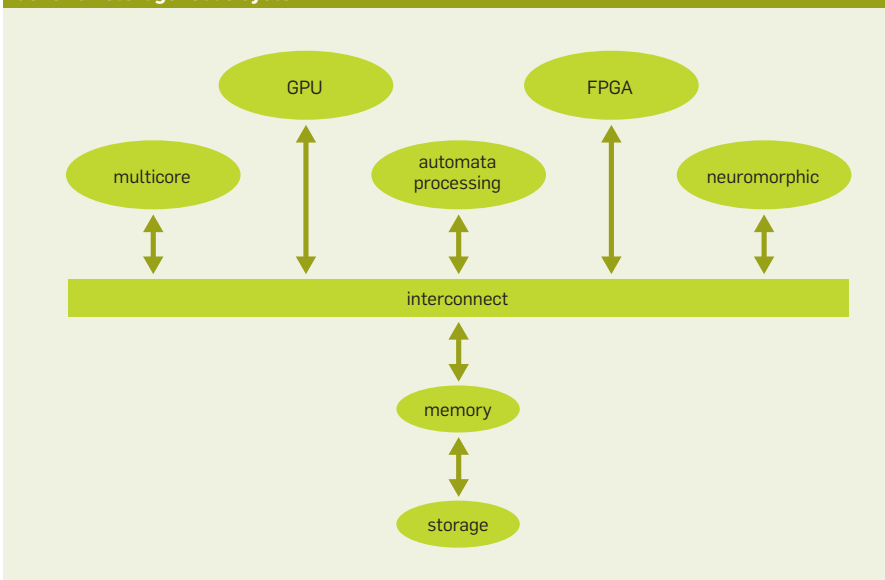
Yet another challenge is distributing the workload among the different cores to get the best performance with the lowest power consumption. The answer to this question must be found across the whole computing stack, from algorithms to process technology.

The move from a single board to multiboard and into high-performance computers also means a move from shared memory to distributed memory. This makes the interconnect and workload distribution even more challenging.

### Software Challenges

At the software level, the situation is also very challenging. How are we going to program these beasts? Sequential programming is hard. Parallel programming is harder. Parallel programming of heterogeneous machines is extremely challenging if we care about performance and power efficiency. There are several considerations: how much hardware to reveal to the programmer, the measures of success,

Generic heterogeneous system.



and the need for a new programming model (or language).

Before trying to answer these questions, we need to discuss the eternal issue of productivity of the programmer vs. performance of the generated software. The common wisdom used to be that many aspects of the hardware needed to be hidden from the programmer to increase productivity. Writing in Python makes you more productive than writing in C, which is more productive than writing in assembly, right? The answer is not that easy, because many Python routines, for example, are just C wrappers. With the proliferation of heterogeneous machines, performance programmers for use by productivity programmers will create more and more libraries. Even productivity programmers, however, need to make some hard decisions: how to decompose the application into threads (or processes) suitable for the hardware at hand (this may require experimenting with different algorithms), and which parts of the program do not require high performance and can be executed in lower-power-consumption mode (for example, parts that require I/O)?

Defining the measures of success poses a number of challenges for both productivity and performance programmers. What are the measures of success of a program written for a heterogeneous machine? Many of these measures have characteristics in common with those of traditional parallel code for homogeneous machines. The first, of course, is performance. How much speedup do you get relative to the sequential version and relative to the parallel version of homogeneous computing?

The second measure is scalability. Does your program scale as more cores are added? Scalability in heterogeneous computing is more complicated than in the homogeneous case. For the latter, you just add more of the same. For heterogeneous machines, you have more options: adding more cores of some type, or more GPUs, or maybe FPGAs. How does the program behave in each case?

The third measure of success is reliability. As transistors get smaller, they become more susceptible to faults, both transient and permanent. Do you leave this issue of dealing with faults to the

hardware, or system software, or shall the programmer have some say? Each strategy has its pros and cons. On the one hand, if it is left to the hardware or the system software, the programmer will be more productive. On the other hand, the programmer is better informed than the system to decide how to achieve graceful degradation in performance if the number of cores decreases as a result of failure or a thread produces the wrong result because of a transient fault. The programmer can have, for example, two versions of the same subroutine: one to be executed on a GPU and the other on several traditional cores.

Portability is another issue. If you are writing a niche program for a well-defined machine, then the first three measures are enough. But if you are writing a program for public use on many different heterogeneous computing machines, then you need to ensure portability. What happens if your code runs on a machine with an FPGA instead of a GPU, for example? This scenario is not unlikely in the near future.

### The Best Strategy

Given these questions and considerations, what is the best strategy? Should we introduce new programming models (and languages), or should we fix/update current ones? Psychology has something to say. The more choices a person has, the better—until some threshold is reached. Beyond that, people become overwhelmed and will stick to whatever language they are using. But we have to be very careful about *fixing* a language. Perl used to be called a “write-only language.” We don’t want to fall into the same trap. Deciding which language to fix/modify is a very difficult decision, and a wrong decision would have a very high cost. For heterogeneous computing, OpenCL (Open Computing Language) seems like a good candidate for shared-memory machines, but it must be more user friendly. How about distributed memory? Is MPI (Message Passing Interface) good enough? Do any of the currently available languages/paradigms consider reliability as a measure of success?

The best scheme seems to be twofold: new paradigms invented and tested in academia while the filtering happens in industry. How does the fil-

tering happen? It happens when an inflection point occurs in the computing world. Examples of two previous inflection points are moving from single core to multicore and the rise of GPUs. We are currently witnessing a couple of inflection points at the same time: getting close to exascale computing and the rise of the Internet of Things. Heterogeneous computing is the enabling technology for both.

Heterogeneous computing is already here, and it will stay. Making the best use of it will require revisiting the whole computing stack. At the algorithmic level, keep in mind that computation is now much cheaper than memory access and data movement. Programming models need to deal with productivity vs. performance. Compilers need to *learn* to use heterogeneous nodes. They have a long way to go, because compilers are not yet as mature in the parallel-computing arena in general as they are in sequential programming. Operating systems must learn new tricks. Computer architects need to decide which nodes to put together to get the most effective machines, how to design the memory hierarchy, and how best to connect all these modules. At the circuit level and the process technology level, we have a long wish list of reliability, power, compatibility, and cost. There are many hanging fruits at all levels of the computing stack, all ready for the picking if we can figure out the thorns. □

### Related articles on queue.acm.org

#### Computing without Processors

Satnam Singh

<http://queue.acm.org/detail.cfm?id=2000516>

#### FPGA Programming for the Masses

David F. Bacon, Rodric Rabbah, and Sunil Shukla

<http://queue.acm.org/detail.cfm?id=2443836>

#### A Conversation with John Hennessy and David Patterson

<http://queue.acm.org/detail.cfm?id=1189286>

### References

1. HSA Foundation; <http://www.hsafoundation.com/>.
2. IBM Research. The cognitive era; <https://www.research.ibm.com/cognitive-computing/>.
3. Micron. Automata processor; <http://www.micronautomata.com/>.

**Mohamed Zahran** is a clinical associate professor of computer science at New York University. His research interests span several areas of computer architecture and hardware/software interaction.

Copyright held by owner/author.  
Publication rights licensed to ACM. \$15.00

Article development led by [acmqueue](http://queue.acm.org)  
queue.acm.org

**Expert-curated guides to  
the best of CS research.**

# Research for Practice: Distributed Transactions and Networks as Physical Sensors

RESEARCH FOR PRACTICE continues in its fourth installment by bringing you a pair of paper selections on distributed transactions and sensing with the aid of physical networks.

First, Irene Zhang delivers a whirlwind tour of recent developments in distributed concurrency control. If you thought distributed transactions were prohibitively expensive, Irene's selections may prompt you to reconsider: The use of atomic clocks, clever replication protocols, and new means of commit ordering all improve performance at scale.

Second, Fadel Adib provides a fascinating look at using computer networks as physical sensors. It turns out that the radio waves passing through our environment and bodies are subtly modulated as they do so. As Fadel's selection shows, new techniques for sensing and interpreting these modulations allow us to perform tasks previously reserved for science fiction: seeing through walls, performing gesture recognition, and monitoring breathing.

As always, we have provided open access to the ACM Digital Library for the relevant citations from these selections so you can dig into and fully appreciate the research papers in each.

During the next several installments of Research for Practice, we will continue our journey through the varied landscape of computer science research areas. In the meantime, we welcome your continued feedback and suggestions for topics. Enjoy!

—Peter Bailis

**Peter Bailis** is assistant professor of computer science at Stanford University. His research in the Future Data Systems group (<http://futuresdata.stanford.edu/>) focuses on the design and implementation of next-generation data-intensive systems.



## **Distributed Transactions By Irene Zhang**

Distributed transactions make it easier for programmers to reason about the correctness of their applications in modern data centers, where both concurrency and failures happen at scale. Distributed storage systems and databases with ACID (atomicity, consistency, isolation, durability) guarantees help programmers by ensuring that opera-

## **>> about RfP**

**Research for Practice combines the resources of the ACM Digital Library, the largest collection of computer science research in the world, with the expertise of the ACM membership. In every RfP column two or more experts share a short, curated selection of papers on a concentrated, practically oriented topic.**



tions from committed transactions are never lost, operations from concurrent transactions do not interleave, and all or none of the operations from a transaction persist, despite failures of application servers or storage servers.

Unfortunately, distributed transactions have long been thought to be prohibitively expensive. In modern storage systems, which partition data for scalability and replicate data for fault tolerance, distributed transactions need coordination at every level: on each storage server, across replicas, and across partitions.

Three recent research papers presented here have made significant strides in reducing the coordination needed for distributed transactions, making them more efficient at every level. The first reduces the cost of read-only transactions across geodistributed data centers using atomic clocks. The second reduces the cost of read-write transactions across replicas by eliminating consistency from the replication protocol. The last reduces the cost of transactions on each storage server using a modular concurrency-control mechanism. Taken together, these papers demonstrate that it is possible to provide distributed transactions with low cost, even at Google scale.

### High-Performance Read-Only Transactions with Atomic Clocks

**Corbett, J. C., et al.**

Spanner: Google's globally distributed database. In *Proceedings of Operating Systems Design and Implementation*, 2012; <http://static.googleusercontent.com/media/research.google.com/en/archive/spanner-osdi2012.pdf>.

Linearizable transactions are useful for programmers because they behave in a way that is easy to understand: there is a single global transaction ordering and it matches the order in which the transactions commit. Unfortunately, linearizable transactions are expensive, especially in a globally distributed system, because they re-

quire every transaction to coordinate with every other transaction, including read-only transactions.

Spanner gets around this problem by using loosely synchronized clocks. Every storage server synchronizes with an atomic clock in the data center, and they estimate the clock skew between servers based on the drift between the atomic clocks. Then Spanner assigns every read-write transaction a timestamp and waits out the clock skew to ensure that the timestamp is in the past, allowing read-only transactions to read at their local current time without any coordination. This technique comes with a caveat, however: if their estimate of the clock skew is off, Spanner no longer guarantees a linearizable transaction ordering.


### High-Performance Read-Write Transactions with Unordered Replication

**Zhang, I., et al.**

Building consistent transactions with inconsistent replication. *Symposium on Operating Systems Principles*, 2015; <https://homes.cs.washington.edu/~arvind/papers/tapir.pdf>.

While Spanner makes read-only transactions less expensive, it does not reduce the cost of read-write transactions. This selection makes the observation that there is wasted work in existing databases when committing transactions: both the transaction protocol and the replication protocol enforce a strong ordering. Thus, it is possible to eliminate the coordination across replicas by using a completely unordered replication protocol and enforce only a linearizable ordering of committed transactions.

The paper introduces an unordered, consensus-based replication protocol, called inconsistent replication, and defines TAPIR (Transactional Application Protocol for Inconsistent Replication) to run on top of it. TAPIR also uses loosely synchronized clocks but as a performance optimization, not a correctness require-



**Three recent research papers have made significant strides in reducing the coordination needed for distributed transactions, making them more efficient at every level.**



**New research bridges state-of-the-art wireless techniques with human-computer interaction. The concepts underlying this new line of research build on basic physical principles of RF waves, such as Wi-Fi.**

ment, avoiding Spanner's caveat. TAPIR represents one option in the design space, but many other possibilities also make for promising lines of research.

### High-Performance Transactions with Modular Concurrency Control

Xie, C., et al.

High-performance ACID via modular concurrency control. *Symposium on Operating Systems Principles*, 2015; <http://sigops.org/sosp/sosp15/current/2015-Monterey/263-xie-online.pdf>.

While much cross-server coordination has been eliminated, transactions can still require significant coordination at each storage server, increasing performance cost. For example, Spanner requires locking, which blocks concurrent transactions that access the same keys, and TAPIR requires optimistic concurrency control, which causes aborts under high contention.

The distributed database system Callas seeks to reduce this cost by grouping transactions based on performance characteristics and applying a concurrency-control mechanism that is best suited for each group. This is made possible through a novel two-tiered concurrency-control mechanism that locks across groups and leaves each one free to use any concurrency-control mechanism, including transaction chopping. The cool thing about the technique is that it can be applied to nondistributed databases as well, although it has the most impact in a distributed system and could probably even be recursively nested for more complex workloads.

### Decreasing Costs Are a Reality

We need to rehabilitate the reputation of distributed transactions. They are powerful tools for application programmers, yet most avoid them because of their perceived cost. While transactional storage will always have a fundamental performance overhead, especially in a distributed environment, these papers show that the overhead need not be exorbitant. Even better, each of these papers points to a promising avenue of research to further reduce the cost of distributed transactions in practical ways, hinting at the possibility that someday

programmers will no longer have to choose between distributed transactions and performance.

Irene Zhang is a Ph.D. student at the University of Washington, where she works in the Computer Systems Lab. Her research focuses on systems for large-scale, distributed applications, including distributed runtime systems and transactional storage systems.



### Networks As Physical Sensors By Fadel Adib

Can Wi-Fi signals allow us to see through walls? For many years,

humans have fantasized about X-ray vision and played with the concept in comic books and sci-fi films. This section highlights recent research that has unlocked the exciting potential of wireless signals and expanded the role of wireless networks, enabling them to deliver new services ranging from seeing through walls to non-contact sensing of heartbeats. To do so, this new research bridges state-of-the-art wireless techniques with human-computer interaction.

The concepts underlying this new line of research build on basic physical principles of RF (radio frequency) waves such as Wi-Fi. Specifically, as these waves travel in the wireless medium, they bounce off different objects—including the human body—before arriving at a receiver; hence, they carry information about the environment. The following selection of papers demonstrates how to extract and analyze this information, allowing wireless networks to be used as physical sensors.

### Seeing Through Walls with Wi-Fi

Adib, F., Katabi, D.

See through walls with Wi-Fi! ACM SIGCOMM, 2013; <http://people.csail.mit.edu/fadel/papers/wivi-paper.pdf>.

The first paper shows that Wi-Fi signals can extend our senses, allowing us to see moving objects through walls and behind closed doors. In particular, such signals can be used to identify the number of people in a closed room and their relative locations. The basic idea is similar to radar and sonar imaging. Specifically, when faced with a nonmetallic wall, a fraction of the wireless signal would

traverse the wall, reflect off objects and humans, and come back imprinted with a signature of what is inside a closed room. To convince yourself that Wi-Fi signals traverse walls, just recall how you can receive Wi-Fi from another room.

The main challenge of using Wi-Fi signals to see through a wall is that the wall's reflection is very powerful. In fact, the wall's reflection is 10,000–100,000 times stronger than any reflection coming from behind the wall. As a result, the wall's reflection will overwhelm the Wi-Fi device and prevent it from detecting any minute reflection coming from behind it. This behavior is analogous to how someone looking at the sun cannot see an airplane in the sky at the same time. The sun's light would overwhelm the person's eyes and prevent them from seeing the airplane, just as the wall's reflection would overwhelm the Wi-Fi receiver and prevent it from detecting reflections from behind it.

To overcome this problem, the authors of this paper leverage recent advances in MIMO (multiple-input, multiple-output) communications. In MIMO, multiple antenna systems can encode their transmissions so that the signal is nulled (that is, sums up to zero) at a particular receive antenna. MIMO systems use this capability to eliminate the interference of unwanted receivers. In contrast, this paper proposes the use of nulling to eliminate reflections from static objects, including the wall. By eliminating the wall's reflection, the proposed system can start registering the minute reflections from behind it. It analyzes these reflections to coarsely track the motion of a person behind a wall and count the number of people in a closed room.

### Gesture Recognition with Wi-Fi

**Pu, Q., Gupta, S., Gollakota, S., Patel, S.**

Whole-home gesture recognition using wireless signals. *ACM MobiCom*, 2013; <https://homes.cs.washington.edu/~gshyam/Papers/wisee.pdf>.

This paper takes Wi-Fi-based motion tracking to another level: it shows how to use Wi-Fi reflections to recognize human gestures. Specifically, over the past few years there has been a growing interest in gesture-based user

interfaces. Past gesture-based interfaces, however, required the person either to be directly in front of a sensor (like the Xbox Kinect) or to wear or carry a device (such as Nintendo Wii). In contrast, this paper shows how to perform gesture recognition throughout an entire home without requiring the user to hold or wear any sensor. It does so by relying on Wi-Fi signals.

To capture information about gestures using wireless signals, this research relies on the Doppler effect. The canonical example of Doppler is the pitch of an ambulance siren that increases as it gets closer and decreases as it moves farther away. The authors leverage this concept using Wi-Fi signals.

In particular, Wi-Fi signals are transmitted at a carrier frequency (around 2.4GHz). A forward movement causes a small increase in this frequency (by a few hertz) and a backward movement causes a small decrease in this frequency. The authors observe that human gestures are typically composed of forward-backward movements. By zooming in on the frequency changes in the reflected signal and decomposing them into small movements, they show how to recognize human gestures. They use this capability to enable users to control appliances throughout their homes by performing in-air gestures.

### Monitoring Breathing and Heart Rate Using Wireless Signals

**Adib, F., Mao, H., Kabelac, Z., Katabi, D., Miller, R.C.**

Smart homes that monitor breathing and heart rate. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, 837–846; <http://witrack.csail.mit.edu/vitalradio/content/vitalradio-paper.pdf>.

The final paper in this selection shows that we can capture and monitor human breathing and heart rates by relying on wireless reflections off the human body. To do so, the authors exploit the fact that wireless signals are affected by any movement in the environment, including chest movements caused by breathing and bodily movements caused by heartbeats.

The main challenge in extracting these minute movements is that they are easily overwhelmed by any other

sources of motion in the environment. To overcome this challenge, the paper first localizes each user in the environment, then zooms in on the signal reflected from each user and analyzes variations in the user's reflection to extract breathing and heart rate. By isolating a user's reflection, it effectively eliminates other sources of interference, including noise or extraneous motion in the environment, which may otherwise mask the minute variations caused by the user's vital signs. This allows multiple users' breathing and heart rates to be monitored using wireless signals, even if the users are behind a wall.

### Where Do We Go from Here?

These papers offer a few instances of a broader set of functionalities that future wireless networks will provide. These networks will likely expand beyond communications and deliver services such as indoor localization, sensing, and control. The papers presented here demonstrate advanced forms of wireless-based sensing to track humans, capture their gestures, and monitor their vital signs even when they do not carry a wireless device. This area of research is still nascent, and only time will tell how much further these techniques can go. ■

**Fadel Adib** is an assistant professor at the MIT Media Lab. He works on wireless networks and sensing systems. His research has been identified as one of the 50 ways MIT has transformed computer science over the past 50 years. The BBC, NBC, CBS, the *Washington Post*, the *Boston Globe*, and *The Guardian* have covered his work.

DOI:10.1145/2993420

**More accessible conferences, digital resources, and ACM SIGs will lead to greater participation by more people with disabilities.**

**BY JONATHAN LAZAR, ELIZABETH F. CHURCHILL, TOVI GROSSMAN, GERRIT VAN DER VEER, PHILIPPE PALANQUE, JOHN “SCOOTER” MORRIS, AND JENNIFER MANKOFF**

# Making the Field of Computing More Inclusive

APPLIED COMPUTER SCIENCE is concerned with the development of algorithms, applications, software, services, methods and measures, and hardware and devices. Excellent work continues to be done to make information technology accessible and usable for people with disabilities. For example, a number of familiar consumer technologies started out designed to provide access to people with disabilities, including the audiobook, speech recognition, captioning, and speech output (screen readers). Speech recognition enables hands-free computing, which is useful in situations like driving. Captioning of videos renders them available to text-based search algorithms but also makes video consumable when ambient sound levels are high, as in airports and gyms. Audiobooks,

which began as a way for blind people to access reading material, are now everyday companions for travelers and commuters everywhere.<sup>9</sup>

In a 2012 *Communications* column, former ACM president Vinton G. Cerf highlighted the importance and difficulty of designing and developing accessible computing systems, making a public call for ideas and reports on success stories and experiences.<sup>5</sup>

Despite the long-term focus on making technology accessible for people with disabilities, the computing profession has not focused on making itself inclusive of people with disabilities; such people remain highly underrepresented at all levels and roles, including practitioner, researcher, student, and teacher.<sup>4</sup> Although the percentage of undergraduate students with disabilities in technology-related majors is fairly representative of the worldwide population as a whole, it is estimated that less than 1% of students who earned Ph.D.'s in computer science (as of 2011) identify as students with disabilities.<sup>13</sup> People with disabilities bring diverse perspectives to the design of technology. Like Cerf, the authors of this article believe becoming more inclusive will be of great benefit to ACM and to technology in general. It is thus important to examine the barriers that exist and determine, as a professional organization, how we can overcome them. This makes strategic and tactical sense; for a professional organization that wants to increase membership, there are many potential community members with disabilities

## » key insights

- **People with disabilities are a potential source of ideas and additional membership for professional computing organizations.**
- **Including people with disabilities in the decision-making processes of professional computing organizations ensures the most important barriers are addressed first.**
- **Processes developed over years are needed to make physical conferences and their related digital content accessible to people with disabilities.**



DOOR IMAGE BY KUTLAVEV DIMITRY WITH CHI PHOTO BY CRYSTAL BUTLER, COURTESY OF CHI 2016 CONFERENCE

who could join the community were it more accessible.

So how do professional organizations in computing start to make themselves more accessible? What needs to be done to enable better access for researchers, practitioners, teachers, and students with disabilities? This article provides an overview of the process and a case study of the steps taken by SIGCHI, the ACM Special Interest Group on Computer-Hu-

man Interaction (<http://www.sigchi.org/>), to be more inclusive for people with disabilities. We note that the term “inclusive” can have a broader meaning that involves economic, geographic, and other types of diversity. In this article, we use the meaning of inclusion found in the fields of education and law, where being inclusive means providing equal opportunity for participation by people with disabilities.

### Addressing Accessibility

SIGCHI is one of ACM’s largest special interest groups, with approximately 3,500 members as of 2016. As with all SIGs, SIGCHI’s core activities are to sponsor conferences, publish articles, and guide and support professional activities through mentoring and career development.

Over the past few years SIGCHI has sought to be more inclusive by decreasing barriers for participation encour-

tered by people with disabilities. In collaboration with other SIGs (such as SIGACCESS, <http://www.sigaccess.org/>), our work has included indirect activities (such as educating conference leadership about disabilities and advocating for inclusion of people with disabilities on committees). We have also improved accessibility at conferences and to digital resources and provided professional-development activities.

We began by recognizing that career development, in all areas of computing, is greatly enhanced through several activities: attendance at conferences on a regular basis; production and consumption of digital resources, from blogs to multimedia content to articles in the ACM Digital Library; and involvement in sponsored mentorship programs. We identified three disability-related concerns that had to be addressed: organization and involvement of stakeholders; considerations regarding physical accessibility; and considerations regarding digital accessibility. Here, we address each in turn. Moreover, we have three corresponding goals in telling the SIGCHI story: underscore the importance of stakeholder engagement; offer broad suggestions for how large SIGs can improve inclusiveness of physical events and digital content; and underscore that addressing physical and digital accessibility is an ongoing process that takes time, with involvement by many stakeholders. The main message is that inclusiveness starts with the creation of an environment of continuous improvement in inclusiveness.

Before discussing them, however, we acknowledge that accessibility is a continuum and SIGCHI (or any other SIG) will not become a highly accessible and inclusive organization overnight.

**Organization of stakeholders.** It is important for the SIGCHI community to have an ongoing process for and platform through which people with disabilities can participate actively. SIGCHI thus created an advocacy group—the SIGCHI Accessibility Community—to work from within SIGCHI to develop best practices for ensuring improved accessibility. It has worked over the past several years on disability-related issues and produced a report<sup>11</sup> documenting accessibility concerns within the SIGCHI community. Jenni-

fer Mankoff, one of the authors of this article, is chair of the SIGCHI Accessibility Community. The other authors are members of the community who have held leadership positions in the SIGCHI Executive Committee or in the conferences, in particular CHI 2014, where many of the practical initiatives were launched and trialed.

**Physical accessibility.** Many people with disabilities report that program committee meetings and conference facilities are often not accessible to people with motor impairments (such as those in wheelchairs). Moreover, elevators are sometimes not available, and few presentation stages have ramps. Processes should thus be planned in advance for requesting disability-related accommodations (such as sign-language interpretation for presentations and easy booking of accessible hotel rooms), and on-site accommodations need to be made available and communicated effectively in promotional materials or websites, as well as at event venues.

**Digital accessibility.** Many computing professionals use the resources available on the central ACM website (such as job banks, blogs, videos, and articles in the ACM Digital Library) that serve as the foundation for information sharing and knowledge growth. Within ACM, each SIG has its own website, with targeted digital resources for the needs of SIG members. Too often, however, the sites and information hosted are not in an accessible format, creating a discriminatory barrier. One approach has been to provide an “information on request” option for people unable to access certain content. But this is not an adequate solution; when digital resources are made accessible only upon request, the amount of material available to someone with a disability is limited and a time delay is introduced. This puts the person with a disability at a disadvantage compared to those without disabilities. Both the delay in time and the limitation in the amount of content available (due to “upon request” accommodations) can be considered forms of discrimination.<sup>9</sup> An informal analysis we conducted at SIGCHI revealed many conference websites, paper-submission processes, and conference-registration processes are not accessible.

## SIGCHI as Case Study

SIGCHI has been addressing accessibility across the areas identified for improvement through a number of experimental initiatives. For example, an accessibility chair was first appointed at SIGCHI’s flagship conference CHI as early as 1996 with some success, but the position did not continue consistently in subsequent conferences. A broader effort was needed, so, in 2011, the SIGCHI Executive Committee began a program to raise awareness and rationalize processes around inclusiveness; see the sidebar “SIGCHI Accessibility Timeline.”

**Education of leadership.** The SIGCHI Executive Committee established a program of information gathering, reaching out to key professional groups and members of the SIGCHI community with disabilities, collaborating explicitly with two groups:

*ACM SIGACCESS.* ACM SIGACCESS is in many ways a role model, with accessible conferences and publications and a large percentage of community members with disabilities. SIGACCESS has documentation and processes for how to make conferences and digital resources accessible for all who want to participate. A core challenge in applying SIGACCESS approaches to the SIGCHI context is the difference in the attendee population. SIGCHI members are not all as aware or committed to accessibility as SIGACCESS members, whose expertise and interest center on accessibility. SIGACCESS also has a longstanding tradition of inclusion, so people with disabilities know their needs will be met at a SIGACCESS conference. SIGCHI needs to build this awareness among its membership, devise inclusive practices, and build a reputation for accessibility. To create awareness, enthusiasm, and engagement within a less-invested membership requires a different set of strategies.

*AccessComputing.* Staff of the AccessComputing project at the University of Washington have been key to SIGCHI’s progress in accessibility. AccessComputing is a National Science Foundation-funded Broadening Participation Alliance that focuses on increasing access to the field of computer science for people with disabilities.<sup>1</sup> At the August 2013 SIGCHI Executive Committee meeting in Seattle, a subgroup

of the SIGCHI Executive Committee working on accessibility met with the AccessComputing leadership team.

Interaction with these groups made clear that a number of recommendations could be made. First, SIGCHI event organizers should be encouraged to appoint accessibility chairs or ensure that an advocate for accessibility would be part of the conference leadership committee. Second, discussions about stakeholder responsibilities should occur to, for example, clarify what aspects of accessibility are under the purview of ACM, vendors (such as website developers), and the conference committee. Such issues could perhaps be resolved or highlighted through appointment of accessibility chairs. Third, SIGCHI should recognize that reliance on volunteers represents a significant barrier to the scalability of accessibility throughout ACM and may be a major factor in limiting what leadership is able to accomplish.

As noted, SIGCHI leadership also discovered how advantageous it is to separate physical accessibility from digital accessibility. Although both are important, rarely in volunteer organizations like SIGCHI do the same people have responsibility for both for several reasons. First, the combination of expertise in physical and digital accessibility rarely resides in one person; for example, it is unlikely a single individual will have great experience in digital document markup languages for accessibility and the guidelines and recommendations for doorframe size and turnaround distance needed for wheelchair accessibility. Second, volunteer time is precious; it can be prohibitively time consuming for one person to take on all such responsibility. During the time period covered here, 2011–2016, within SIGCHI, the vice president (VP) of conferences and the general conference chairs for each sponsored and in-cooperation conference would have responsibility for physical accessibility. For digital accessibility, the VP of operations (for the website), the VP of publications (for the content being published), and the conference technical program chairs would have responsibility. The SIGCHI Executive Committee created a new structure—the CHI Steering Committee—in 2016 to oversee the activities of all conference

## SIGCHI Accessibility Timeline

The following is a timeline of SIGCHI's actions related to accessibility:

**2011.** Focused discussions on accessibility and inclusiveness begin at SIGCHI Executive Committee meetings.

**2012.** The SIGCHI Conference Management Committee begins using the SIGACCESS conference checklist at on-site facility walkthroughs; note it affected only locations that, at the time, were not yet contracted though is now in place for all future conferences.

**2013.** The Executive Committee creates a formal plan for inclusiveness at its spring meeting.

*Email alias.* An email alias is created to invite SIGCHI members to share accessibility suggestions and provide a way for them to report problems;

*Inclusiveness.* The issue of inclusiveness is raised by the Executive Committee at the CHI 2013 Town Hall meeting in Paris;

*Questions.* Questions about accessibility and inclusiveness are added to the CHI 2013 post-conference survey and to all subsequent CHI post-conference surveys;

*Accessibility chairs.* The positions of “digital accessibility chair” and “physical accessibility chair” are added to the CHI 2014 committee;

*AccessComputing.* The Executive Committee meets with AccessComputing directors at the Executive Committee's summer meeting;

*Papers.* The webpage labeled “Information about making your CHI paper accessible” is added to the CHI 2014 conference website;

*Website and app.* Two experts evaluate the CHI 2014 website and related mobile app for accessibility;

*Accommodations.* Questions about disability-related accommodations are added to CHI 2014 registration forms and to all subsequent CHI registration forms;

*Automated reports.* All authors of accepted papers for CHI 2014 receive an automated report evaluating the accessibility of their submissions; and

*Accessibility Community.* The SIGCHI Accessibility Community is created.<sup>10</sup>

**2014.** First face-to-face meeting of the SIGCHI Accessibility Community is held at the CHI 2014 conference in Toronto.

*Chairs appointed.* Digital accessibility chairs and physical accessibility chairs are appointed to the CHI 2015 Technical Program Committee;

*Discussions.* Inclusiveness is discussed at the CHI 2014 Town Hall meeting in Toronto; and

*Officers elected.* For the first time, officers for the SIGCHI Accessibility Community are elected.

**2015.** The first report examining SIGCHI accessibility is produced, documenting failures and successes of CHI (and SIGCHI-sponsored) conferences to meet the accessibility needs of attendees.

**2016.** The SIGCHI Executive Committee authorizes use of SIGCHI funds to create closed captions for all videos on the SIGCHI YouTube channel (<https://www.youtube.com/user/acmsigchi>).

*Telepresence robots.* Individuals with disabilities unable to travel were encouraged to apply for the use of telepresence robots (deemed a success) at CHI 2016.

*Appointed.* Individual appointed to CHI Steering Committee to specifically work on accessibility.

committee chairs from CHI 2018 onward. One of the co-authors of this article, Jennifer Mankoff, was appointed to the steering committee to supervise implementation of a consistent level of accessibility throughout all SIGCHI-sponsored conferences.

### SIGCHI Accessibility Community

As more feedback and suggestions became available, it was necessary to prioritize requests in light of limited resources. Meeting in August 2013, the SIGCHI Executive Committee decided to crowdsource some of the feedback and priority setting. There is a mecha-

nism on the SIGCHI website for the formation of SIGCHI “communities” in which members with a similar interest are able to use certain features on the website, including voting and resource sharing.<sup>12</sup> At the same meeting of the Executive Committee, several people who had been involved in the discussions about improving SIGCHI accessibility were invited to form a SIGCHI community on the topic of accessibility. Unlike SIGACCESS, the SIGCHI Accessibility Community's primary functions are to provide feedback to SIGCHI on accessibility efforts, help set priorities, and provide the op-

portunity for people with disabilities or those who are committed to improving accessibility to advance such efforts. The first face-to-face meeting of the SIGCHI Accessibility Community was held at the CHI 2014 conference in Toronto and its first officers were elected in November 2014. Today, it lists 53 official members on the SIGCHI website and 134 members in the Facebook interest group.

The mission of the SIGCHI Accessibility Community, as spelled out on the website, is to improve “... the accessibility of SIGCHI conferences, and the digital accessibility of SIGCHI web site and publications. Our priorities include providing clear support and information to conferences and their leadership about accessibility, providing support for SIGCHI members who are facing accessibility issues, advocating for accessibility issues, and liaising with other communities such as SIGACCESS.” One of the first acts of the SIGCHI Accessibility Community in 2014 was to assess the state of accessibility across SIGCHI from a member perspective, conducting a survey of SIGCHI members and analyzing post-conference survey responses given by CHI attendees about CHI accessibility. Other data analyzed included the number of conferences in 2014 sponsored by SIGCHI with accessibility chairs (four of 17) and reports by community members on problems they had encountered. This led to the SIGCHI Accessibility Community’s May 2016 report,<sup>11</sup> including five recommendations for future goals for SIGCHI:

*Recommendation 1.* Ensure 100% of conferences are accessible, have an accessibility policy, and have a clear chain of command for addressing accessibility issues;

*Recommendation 2.* Ensure 100% of new content (such as videos and papers) meets established standards for accessibility and develop a process for achieving this goal;

*Recommendation 3.* Create a process for handling accessibility requests within SIGCHI;

*Recommendation 4.* Increase representation of people with disabilities within SIGCHI; and

*Recommendation 5.* Assess SIGCHI’s success in meeting accessibility guidelines at least once every two years.



## The main message is that inclusiveness starts with the creation of an environment of continuous improvement in inclusiveness.



The SIGCHI Accessibility Community brought one major concern—accessibility of other SIGCHI-sponsored conferences—to the attention of the Executive Committee: Although the flagship CHI conference is steadily improving accessibility, most other SIGCHI-sponsored or in-cooperation conferences have taken no steps toward improving accessibility. The Accessibility Community has also highlighted key factors affecting accessibility that need to be addressed, including lack of a clear process (from the member perspective) for handling accessibility problems and constraints; the burden of negotiating accessibility on a case-by-case basis; the problems of depending entirely on volunteers to assess and improve accessibility; and the lack of accessibility at venues (such as in program committee meetings).

**Physical accessibility.** SIGCHI efforts related to physical accessibility have been evolving for several years. The SIGCHI Conference Management Committee first adopted the SIGACCESS conference physical-accessibility checklist for meeting and conference-site walkthroughs in 2012.<sup>a</sup> The first direct engagement with membership as a whole about physical accessibility was at the CHI 2013 conference in Paris, where SIGCHI leadership heard complaints about the venue’s lack of physical accessibility. Discussion at the SIG Town Hall meeting at the conference led to adding a post-conference survey question regarding physical accessibility, resulting in 29 responses. Four issues were cited, the first two relating to hotel accommodations and the third and fourth to the convention venue itself:

*Closest hotel.* The closest recommended hotel was inaccessible for those using a wheelchair or scooter;

*Connecting paths.* Supposedly accessible connecting paths between the hotels and the convention center were poorly signed and not consistently open;

*Ramps.* At the convention center, presenters needing wheelchair or scooter access could not easily reach

<sup>a</sup> Because conference venues are contracted years in advance, walkthroughs in 2012 affected only conferences held in 2015 and later; for the checklists, including the “accessible conference guide,” see <http://www.sigaccess.org>



stages, requiring portable ramps to be added; and

*Distance.* The vast size of the convention center meant considerable distance between events, affecting attendees with mobility limitations.

Based on the data collected, SIGCHI leadership concluded that two categories of data or communication were missing between organizers and attendees for the organization's conferences:

*Attendees.* Attendees, especially presenters, need a mechanism for letting conference planners know in advance if they require any type of special accommodations; and

*Conferences.* Conferences need to let potential attendees know in advance which meeting locations and hotel accommodations are accessible and which are not and provide specific directions (and, where appropriate, signage) to guide attendees along accessible routes between hotels and convention centers.

To address the first, a box was added to the subsequent conference registration form for CHI 2014, as well as for 2015 and 2016. The online forms invite authors of accepted papers/notes to indicate if the presenters of the papers/notes will need any type of disability-related accommodation and, if so, what type; for example, SIGCHI indicated it would fund as many sign-language interpreters as needed, but they must be requested in advance. To address information flow, a webpage was set up for the CHI 2014 conference website by the conference management team, the chairs, and the SIGCHI executive VP dedicated to physical accessibility, including detailed information regarding transportation and convention center and hotel contacts. The same information was provided for the CHI 2015 and CHI 2016 conferences. In addition, the committee in charge of venue selection began (as discussed in the sidebar's timeline) to assess site accessibility so a basic level of access can be ensured (such as wheelchairs and scooters being able to get to every part of the conference).

In 2014, SIGCHI leadership continued to ask about accessibility in the post-conference survey; while such survey data is not public, summaries of the data are included in reports from the SIGCHI Accessibility Community.<sup>11</sup> From

the survey, 623 CHI 2014 attendees answered the question about accessibility, with only 12 indicating their expressed needs were not met and the rest that their needs were met. Only one of those 12 responses actually indicated a specific disability-related need that was requested but not met. The other responses indicated an accommodation that should have been requested but was not ("I had an accessibility-related special need but did not request an accommodation"); most of the comments related to the cost of the conference or labeling of food ingredients. Although these topics relate to the inclusiveness of the conference, none specifically related to perceptual, motor, or cognitive disabilities. In addition, one change has been made though not based on the feedback from surveys; several related conferences (such as ASSETS and ubiComp) allow telepresence robots (such as Beam from Sutable Technologies, Inc. of Palo Alto, CA) to allow for participation of individuals with disabilities who are unable to travel. The CHI 2016 conference committee accepted applications from members who wanted to participate in the conference via a Beam robot due to "mobility impairments, chronic health issues, or temporary travel limitations." The experiment with robots at CHI 2016 was deemed a success, with a total of 35 individuals participating via 10 telepresence robots.

**Digital accessibility.** For the CHI 2014 digital accessibility chair, three

topical areas were suggested by the conference chairs for improvement: conference website, conference mobile apps, and papers-publication process.

Among them, the most challenging was the papers review process. There is one clear international technical standard for webpages—the Web Content Accessibility Guidelines (WCAG) version 2.0—that has been adopted by many national governments, educational organizations, and corporations.<sup>14</sup> The guidelines were used in May 2013 in two preliminary evaluations of accessibility—one by a SIGCHI Executive Committee member and one by the AccessComputing Project at the University of Washington—and changes were made to the website (minor tagging of images) to improve accessibility. This was a good starting point but not optimal because there should be more evaluations involving people with disabilities. A similar process was used for the CHI 2015 and the CHI 2016 conferences, and it is hoped the SIGCHI Accessibility Community can be involved in the future to perform user-based accessibility evaluations.

The technical program chair and digital accessibility chair for CHI 2014 learned that the papers-publishing company SIGCHI works with, Sheridan, offers the option of evaluating accepted-paper .pdf files for accessibility and notifying authors of violations. However, this option was not possible



CHI16 telepresence robots at recharging station.

for the CHI 2014 conference because the timeline and contract with the company had already been fixed. It will thus be investigated for future conferences for which contracts have not been set; the CHI 2015 contracts had already been signed, and the CHI 2016 committee decided not to take the option.

Many guides to .pdf accessibility assume much knowledge about .pdf design and provide a high level of detail about every possible violation. Unlike the WCAG 2.0 for webpages, there is no one clear, agreed-upon standard for .pdf documents. From all the various guidelines, from SIGACCESS and the various international standards bodies, the CHI 2014 papers review committee eventually adopted five recommendations for implementation for the CHI 2014 papers, in consultation with the AccessComputing group. The information was provided to authors on the conference website,<sup>6</sup> and the same guidelines were used for CHI 2015 and CHI 2016. The focus was on improving aspects of .pdf accessibility specifically related to CHI papers, including alternative text provided for images, table headers, generating a tagged .pdf, default language information in the .pdf, and having a correct tab order; readers are encouraged to visit the guide<sup>6</sup> for more on these recommendations. A detailed guide was created to provide step-by-step instructions for the five main recommendations. The goal was to maximize accessibility while minimizing the workload of individual authors.

Information on .pdf accessibility, including a step-by-step guide for adding accessibility information and tool information on checking a .pdf, was added to the CHI 2014 website, and information about .pdf accessibility was added to the CHI 2014 paper tem-

plates. This same information was used for CHI 2015 and CHI 2016.

The CHI 2014 conference received 2,043 submissions for papers and notes, with 465 accepted for publication. For all 465, the CHI 2014 team ran an automated check using Adobe Acrobat Action Wizard to create an accessibility report for each submission, creating a spreadsheet identifying which of the five recommendations each submission had addressed. The papers review committee sent a report to the primary authors on their submission's accessibility features, including links to the instructions for each of the recommendations. Authors received it before the camera-ready copy was to be submitted and were reminded to make their papers compliant with the five recommendations. The goal was to inform, educate, and improve digital accessibility. Making the .pdf file accessible was thus encouraged but not required. This action increased accessibility of accepted papers that were published in the ACM Digital Library but did not increase accessibility of the paper reviewing process. Furthermore, there are challenges with using some of the existing document production tools to create accessible .pdf files. Not all of the commonly used word processors and text editors support making accessible .pdf files; for example, MS-Word for Mac does not. In addition, although some previous attempts had sought to improve accessibility for LaTeX (such as Babett Schalit's accessibility package<sup>8</sup>), those packages were not robust enough for general use for CHI 2014 and CHI 2015. Nevertheless, SIGCHI volunteers have continued to improve the group's LaTeX templates (such as LaTeX Accessibility<sup>8</sup>) and encourage participation by interested accessibility researchers and SIGCHI au-

thors. In addition, SIGCHI maintains an up-to-date wiki page describing current best practice for creating accessible .pdf documents.<sup>2</sup>

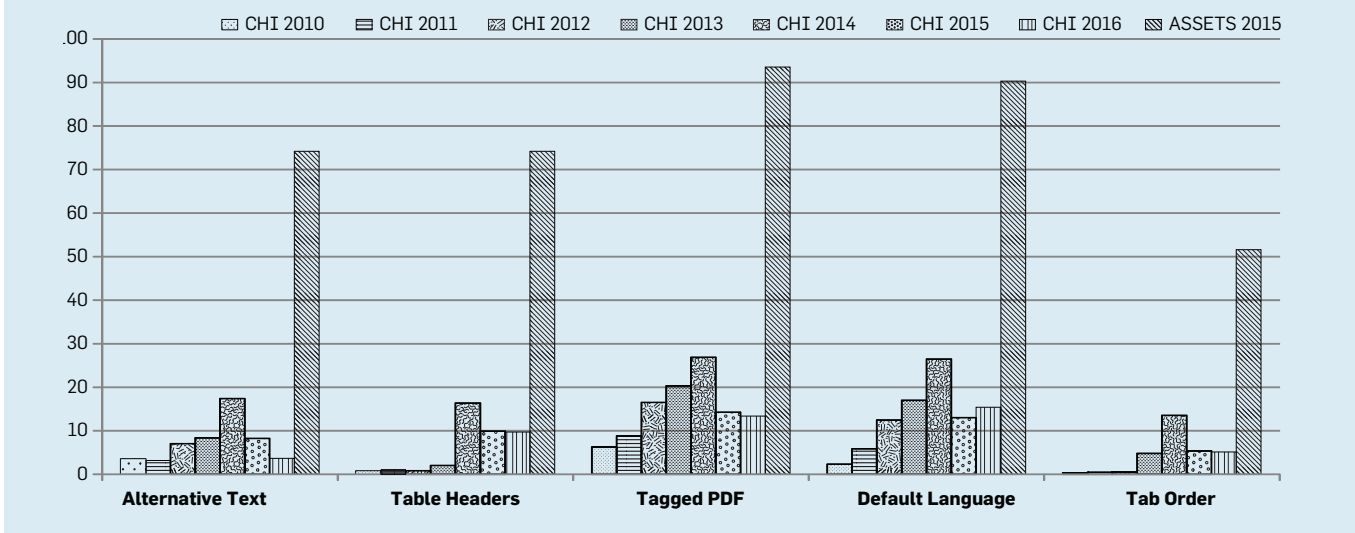
Unfortunately, the CHI 2015 and CHI 2016 conferences did not use the same approach as was used for CHI 2014 of providing specific feedback to authors on the accessibility of their papers. However, the CHI 2015 digital accessibility chair offered to have his research group (from Carnegie-Mellon University) make papers accessible for authors, with 25 authors requesting the service.<sup>3</sup> Although the service was not continued in 2016, the instructions on accessibility and the information in the paper template were still included. While the goal should be year-to-year consistency, having different approaches tested each year does give some useful data for future conference chairs.

Based on data collected by the CHI 2014 Conference Committee, accessibility of papers at CHI 2014 improved compared with previous years. Figure 1 shows the percentage of published CHI papers from 2010 to 2016 that included each of the five recommended accessibility features. The data in the figure indicates the accessibility reports sent to authors in 2014 helped encourage accessibility of papers. The accompanying table lists the same data from the figure in tabular form, showing compliance in four of five (not tab order) categories rose from 16% to 26%, much higher than in previous years. In every category of accessibility feature, the papers submitted were more accessible in 2014 than in any previous year of the CHI conference, though they were not 100% accessible, which is indeed the goal. A separate analysis confirmed that the accessibility of CHI papers improved in 2014.<sup>3</sup> However, without giving the authors individual notification of their papers' accessibility between acceptance and camera-ready submission in 2015, the accessibility levels of papers dropped between 2014 and 2015. Averaged over the five measures of accessibility, the accessibility of papers between 2014 and 2015 dropped nearly 50%. Figures were generally consistent between 2015 and 2016, except for the alternative text, which dropped by more than 50%, with 8.26% compliant in 2015 compared to 3.67% compliant in 2016.

**Percentage of published papers that adhered to each of the five recommendations (%), 2010–2016.**

	Published CHI Papers (% following the guidelines)						
	CHI 2010	CHI 2011	CHI 2012	CHI 2013	CHI 2014	CHI 2015	CHI 2016
Alternative Text	3.6	3.2	7.0	8.4	17.4	8.3	3.7
Table Headers	0.7	1.0	0.8	2.0	16.3	9.9	9.7
Tagged PDF	6.3	8.8	16.5	20.3	26.9	14.3	13.4
Default Language	2.3	5.9	12.5	17.0	26.5	13.0	15.4
Tab Order	0.3	0.5	0.5	4.8	13.5	5.4	5.1

**Figure 1. Percentage of published papers that adhered to each of the five recommendations (%), 2010–2016. The bars here and in Figure 2 are covered in patterned fill, rather than colors, to make the graphs more inclusive for colorblind readers.**



Giving authors individual notification of their papers’ accessibility between acceptance and camera-ready submission in 2014 clearly increased the level of accessibility compliance. While accessibility of papers did increase, 16% to 26% is still not ideal, with a long way to go. As a comparison, we analyzed the accessibility of published papers from the ASSETS 2015 conference, though the sample size for ASSETS papers was 31, much smaller than the number of CHI papers in any given year. ASSETS generally uses two different approaches that have not yet been attempted by the CHI conference: The first is that authors are required (not just encouraged) to make their papers accessible and the second that SIGACCESS, sponsor of the ASSETS conference, specifically requires the company that is contracted for publishing, Sheridan, to manage the accessibility process and check for accessibility. We do not know the specifics of what is required in its contracts with Sheridan, and it is possible Sheridan is required to check for different accessibility features than in our evaluation. Given identical criteria, compliance for ASSETS 2015 papers was much higher than for CHI papers (in any given year) but still not at the 100% goal. In 2015, 74.1% of the ASSETS papers had alternative text and table headers, 93.5% had generated a tagged .pdf file, and 90.3% had default-language information included in the .pdf, but only 51.6% of ASSETS 2015 papers had a correct tab order.

**Figure 2. Difference in adherence among the 465 accepted papers for CHI 2014 between submitted and final versions (%). The bars here are likewise covered in patterned fill, rather than colors, to make the graphs more inclusive for colorblind readers.**

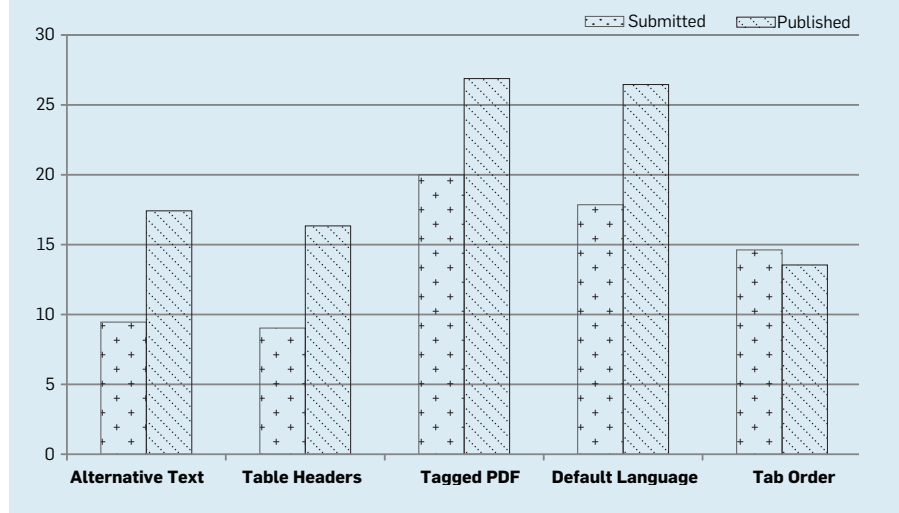


Figure 2 shows the difference in adherence between initial and final submissions for the 465 papers accepted for publication at CHI 2014, where authors were given specific details on the accessibility barriers of their respective papers. In four of five recommendations, accessibility of the papers increased 5% to 10% based on authors receiving feedback on accessibility. Unclear is why adherence to one recommendation (tab order) decreased slightly. There may be cases where authors had to update their final submission based on feedback from the publication vendor and forgot to reapply the accessibility changes.

Note that 30% accessibility of published papers or even 60% accessibil-

ity is not ideal. The goal, as spelled out by the SIGCHI Accessibility Community, is 100% compliance. However, accessibility is a multi-pronged effort, and paper accessibility gets attention because it is an easy-to-measure metric; equally important are many other details we have discussed here (such as having accessibility chairs at each conference, proper information flows, and accessible physical locations). For instance, in choosing the site for the CHI 2019 conference—Glasgow, U.K.—accessibility criteria were specifically taken from the city’s proposals, as well as from on-site walkthroughs, which led to one city with a fully accessible conference venue being chosen

over another equally attractive city but with a conference venue with multiple accessibility barriers.

In addition to event-specific efforts (such as those described here), other efforts to improve accessibility have been ongoing on multiple fronts within SIGCHI over the past few years. For instance, a SIGCHI email alias—sigchi-accessibility@listserv.acm.org—was set up for members to share their concerns with the Executive Committee, underscoring SIGCHI's commitment to being open and welcoming to academics, researchers, and practitioners with disabilities by inviting comments and concerns related to the organization's websites, publications, or physical accessibility at any SIGCHI-sponsored events, including conferences.

Another example of progress involves video captioning. SIGCHI captures the video and slides of a selection of the presentations at CHI and other SIGCHI-sponsored conferences. These presentations are included with the .pdf of the papers in the ACM Digital Library. Starting in 2016, SIGCHI volunteers began to work with ACM to create an ACM SIGCHI YouTube channel to host much of this content. As part of the effort, the SIGCHI Executive Committee authorized use of SIGCHI funds to create closed captions for all the videos on its YouTube channel. Once a video is uploaded to YouTube, SIGCHI works with a captioning company to develop professional (not automated) captioning. Because the captions are human generated, the time to caption all the videos in a conference can vary depending on the total number of videos uploaded.

### **Suggestions for All Computing Organizations**

SIGCHI members surveyed as part of the SIGCHI Accessibility Community Report<sup>11</sup> were typically not aware of any SIG or ACM policy or procedure regarding inclusiveness for people with disabilities. This was the case for those with and those without disabilities. For example, respondents reported<sup>11</sup> being unable to answer the following questions: How can someone with a disability participate in a mentorship program sponsored by the organization? What happens when someone who is blind wants to vote in an election or

run for office? Are the online tools utilized by journal editorial boards accessible? Do the procurement processes for these large contracts include accessibility? And what policies are used for remote participation?

Based on the SIGCHI experience, we can say that professional organization inclusiveness begins with explicit discussions on inclusiveness, and awareness and discussion represent an important first step. Executive committees of SIGs should start the discussion, which should expand to include conference chairs. Conference chairs should discuss accessibility with their technical program chairs. Executive committees should contact members of the professional community with known disabilities and email distribution messages asking for input and feedback. Conference chairs should also be aware that some disabilities are “invisible disabilities” that might not be apparent (such as learning disabilities and disabilities affecting energy level, as with Lupus and Lyme Disease). Starting the discussion produces information sharing, which should lead to a more formalized structure like a policy or specific committee position (such as accessibility chair for a conference). None of these changes will happen overnight. Becoming more inclusive is a process that takes place over a period of years. We thus recommend the following six actions for all ACM SIGs:

*Reach out to SIGACCESS.* No one within ACM has more experience with accessibility issues than SIGACCESS. At various points, SIGCHI used the SIGACCESS conference accessibility guidelines and portions of the SIGACCESS document accessibility guidelines and consulted with various members of the SIGACCESS Executive Committee who were always happy to help. It may be the SIGACCESS solutions cannot be implemented directly by another SIG due to scalability or lack of expertise, but SIGACCESS has the experience of creating solutions for most accessibility issues. SIGACCESS officers welcome inquiries and contacts from other SIGs.

*Encourage proactive involvement and foster bidirectional communication.* Make it easy for community members to notify the organization of potential accessibility needs before events like

conferences, thus allowing appropriate accommodations to be made; if such accommodations are not possible, individuals can be warned. SIG and conference organizers must be clear and up front about accessibility at a conference, answering: What, from a physical point of view, is accessible, and what is not? What barriers will attendees face? And is there a hotel that may be farther away but that involves fewer barriers? Encourage feedback from the community at events and between events.

*Include people with disabilities in organizational processes.* One of the mottos of the disability rights movement is “nothing for us without us.” Decisions about accessibility need to be made based on feedback from those with the most experience—people with disabilities. It is important early on to identify members of your community with disabilities who can provide specific feedback. Acknowledge that perspectives may be skewed; if your community includes many people with one type of disability, the feedback you receive may be biased. A core advisory group can provide feedback and advice and can help determine priorities.

*Be clear about your priorities and communicate rationales.* It is important to acknowledge that everything cannot be done at once. For instance, for an organization starting to become more inclusive, which of the following is a better first step: Making papers accessible or making videos on the website accessible? Making mentorship programs more inclusive or making journal editorial board software more accessible? Making the conference facility selections more accessible or setting up programs for remote attendance? All are important goals that should be achieved over time, but all cannot be achieved immediately. A dedicated advisory group, as with SIGCHI's Accessibility Community, can be useful in setting priorities. Once priorities are set, they need to be communicated to the membership and to the broader community.

*Recognize and explicitly address and communicate trade-offs.* Be open about the fact that there are often trade-offs, as in the one between internationalization and consistent models of accessibility. Part of being an international organization means holding confer-

ences all over the world, including locations that have different accessibility requirements and accommodations. Such trade-offs should be acknowledged. When practices differ, it is critical that they be explicitly documented and communicated.

*Allocate budget from SIG funds.* Allocate budget from your SIG funds to support professional services (such as video captioning). Be clear about what work is done by volunteers and what is outsourced to professional services. SIGCHI and ACM function primarily through their volunteers, but SIGCHI has decided some aspects of accessibility are so important that we must contract with professionals who can provide dedicated and reliable focus to drive our inclusiveness agenda forward. This is not a criticism of the volunteers; all are committed to these initiatives, but for many, such plans are not their primary work focus, so a reliable, accountable effort is not a reasonable expectation.

**Conclusion**

We have three goals in telling the SIGCHI story: underscore the importance of stakeholder engagement; offer broad suggestions for how large SIGs can improve the inclusiveness of physical events and digital content; and underscore that addressing physical and digital accessibility is an ongoing process that takes time, with involvement of many stakeholders. These stakeholders must work together to drive the creation of acceptable and accepted guidelines and resources, find individuals with expertise to work in an advisory capacity, and find volunteers to implement effective strategies and provide feedback regarding the policies and guidelines in action.

Improving the inclusiveness of any organization is a long-term process. It involves planning, structure, and information sharing. It involves checklists and inspections. It involves a commitment to programmatically raising awareness through communication and action. But where does inclusiveness start? One possibility is with members of the specific community raising awareness about barriers. But we advocate a more proactive stance. A professional community that has not been inclusive of people with disabili-

ties is not likely to have members with disabilities who will raise awareness of what is needed. Inclusiveness must start with proactive outreach to increase inclusiveness so change can be driven from within the organization. A reactive stance through which accessibility issues are dealt with as (and only if) they occur is not programmatic and will not be as effective.

The impact of greater accessibility can be profound. The more accessible an organization becomes, the more people will feel comfortable giving feedback and working actively toward inclusive solutions that can lead to more members. As Kirkham<sup>7</sup> said about the current situation, “In practice significantly more research is being done about people with disabilities than by people with disabilities within SIGCHI.” SIGCHI’s hope is that SIGCHI will be a community that is perceived as welcoming for all researchers and practitioners with disabilities.

In addition, actions on the part of any organization, including a SIG community, have the ability to influence outside actors. Large SIGs, when they educate others about digital and physical accessibility, can have significant influence on the conference locations they rent and the universities and companies that employ their members.

ACM has a leading role to play by ensuring all SIGs strive to be inclusive and by thus being a role model for other professional associations. The best way to handle such responsibility would ultimately be to ensure there are professional staff supporting and centralizing the most vital accessibility needs and accessibility is included in contractual relationships (such as with organizations that produce ACM’s website and publications and contract conference venues).

**Acknowledgments**

The authors would like to acknowledge the advice provided by Sheryl Burgstahler, Richard Ladner, Clayton Lewis, Jennifer Rode, Terry Thompson, Shari Trewin, and Jeff Bigham to the SIGCHI Executive Committee and CHI 2014 and CHI 2015 Program Committees. □

**References**

1. AccessComputing: The Alliance for Access to Computing Careers; <http://www.washington.edu/accesscomputing/>

2. Accessibility; <https://github.com/sigchi/Document-Formats/wiki/Accessibility>

3. Brady, E., Zhong, Y., and Bigham, J. Creating accessible PDFs for conference proceedings. In *Proceedings of the 12th Web for All Conference* (Florence, Italy, May 18–22). ACM Press, New York, 2015, article 34.

4. Burgstahler, S., Ladner, R., and Bellman, S. Strategies for increasing the participation in computing of students with disabilities. *ACM Inroads* 3, 4 (2012), 42–48.

5. Cerf, V. Why is accessibility so hard? *Commun. ACM* 55, 11 (Nov. 2012), 7.

6. CHI 2014 Program Committee. CHI 2014 Guide to an Accessible Submission; <http://chi2014.acm.org/authors/guide-to-an-accessible-submission>

7. Kirkham, R., Vines, J., and Olivier, P. Being reasonable: A manifesto for improving the inclusion of disabled people in SIGCHI conferences. In *Extended Abstracts of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea, Apr. 18–23). ACM Press, New York, 2015, 601–612.

8. LaTeX Accessibility; <https://github.com/sigchi/Document-Formats/issues/1>

9. Lazar, J., Goldstein, D., and Taylor, A. *Ensuring Digital Accessibility Through Policy and Process*. Elsevier/Morgan Kaufmann Publishers, Amsterdam, the Netherlands, 2015.

10. SIGCHI Accessibility Community; <http://www.sigchi.org/communities/access>

11. SIGCHI Accessibility Report; <https://docs.google.com/document/d/1XBUV9kl55D430wzJf70d4EvQWR7AIFj7nQQ3z2X26bjw/edit?pref=2&pli=1#heading=h.w0fjbae5fw42>

12. SIGCHI Communities; <http://www.sigchi.org/communities>

13. Taylor, V. and Ladner, R. Data trends on minorities and people with disabilities in computing. *Commun. ACM* 54, 12 (Dec. 2011), 34–37.

14. World Wide Web Consortium (Web Accessibility Initiative). *Web Content Accessibility Guidelines 2.0*; <https://www.w3.org/TR/WCAG20/>

**Jonathan Lazar** (jlazar@towson.edu) is a professor of computer and information sciences and director of the Undergraduate Program in Information Systems at Towson University, Towson, MD, and recipient of the SIGCHI 2016 Social Impact Award.

**Elizabeth Churchill** (churchill@acm.org) is a director of user experience at Google, San Francisco, CA, and Secretary/Treasurer of ACM.

**Tovi Grossman** (tovi.grossman@autodesk.com) is a distinguished research scientist in the User Interface Research Group at Autodesk Research, Toronto, Canada.

**Gerrit C. van der Veer** (gerrit@acm.org) is an emeritus professor of multimedia and culture at the Vrije Universiteit Amsterdam, the Netherlands, guest professor of human-media interaction at Twente University, Twente, the Netherlands, of human-computer and society at the Dutch Open University, Heerten, Netherlands, of interaction design at the Dalian Maritime University, Dalian, China, and of animation and multimedia at the Lushan Academy of Fine Arts, Shenyang, China.

**Philippe Palanque** (palanque@irit.fr) is a professor of computer science at Université Paul Sabatier Paul Sabatier – Toulouse III, France, and head of the Interactive Critical Systems research group of the IRIT laboratory, Toulouse, France.

**John “Scooter” Morris** (scooter@cgl.ucsf.edu) is an adjunct professor in the Department of Pharmaceutical Chemistry at the University of California San Francisco and executive director of the Resource for Biocomputing, Visualization and Informatics, a U.S. National Institutes of Health Biomedical Technology Research Resource at the University of California San Francisco.

**Jennifer Mankoff** (mankoff@cs.cmu.edu) is a professor in the Human Computer Interaction Institute at Carnegie Mellon University, Pittsburgh, PA.

© 2017 ACM 0001-0782/17/03 \$15.00



Watch the authors discuss their work in this exclusive *Communications* video. <http://cacm.acm.org/videos/making-the-field-of-computing-more-inclusive>

DOI:10.1145/2959086

**Along the way, acquire technical expertise and a master's degree, even while changing positions and companies.**

BY DANIEL J. MAZZOLA, ROBERT D. ST. LOUIS, AND MOHAN R. TANNIRU

## The Path to the Top: Insights from Career Histories of Top CIOs

WHEN SEARCHING FOR IT talent, the position of chief information officer (CIO) may be the most difficult to fill successfully. The impact of IT on business value and organizational performance has been extensively discussed in both the academic literature<sup>2,11</sup> and the practitioner literature.<sup>5</sup> All the findings point to the important role the CIO plays in the success of the overall business. This makes it important to understand the traits and characteristics effective CIOs share and the educational and workplace experiences that increase their likelihood of attaining and retaining the CIO mantle, so organizations may be able to identify and groom high-potential CIO candidates and provide career advice to aspiring CIO candidates.

In the early 1980s, an in-depth look by Tanniru<sup>14</sup> at positions held by IT managers before they reached

their first leadership role identified two primary career paths: business and technical. A programmer or analyst entry position led to either a business analyst or technical specialist role. Each such role led to either an IT leadership position or a technical manager position. The past three decades have dramatically changed both the IT and the business landscapes. This is an exploratory follow-up to that study. The goal here, as it was then, is to track the career paths of senior IT leaders—CIOs—and use that information to guide the skill development and career progression of today's IT talent.

More specifically, the research objective of this article is to identify the defining career experiences and educational characteristics of the rungs of the CIO ladder to provide insight for both the firms that hire CIOs and the IT professionals who aspire to be CIOs. The career histories of many CIOs can be discovered through social media data and is the source of the data in this study. We used an inductive methodology to analyze these histories in order to elicit the key identifying features of IT workers who move up the CIO ladder. We categorized the raw data into industry and job types in order to develop a framework that captures key insights and themes that can be used to guide the actions of aspiring CIOs and the firms that recruit them. These initial results suggest an approach for helping workers with the potential for IT leadership to achieve that potential. We conclude with a discussion of future research possibilities for building on these exploratory results.

### Research Methodology

Unlike the Tanniru study in the 1980s, when data was collected from a conve-

#### >> key insights

- Since the mid-1970s, the time required to become a CIO has decreased significantly.
- The institution one graduates from has no influence on one's chances of becoming a CIO.
- Many CIOs have no experience beyond IT.



IMAGE BY LIGHTSPRING

nience sample—known IT professionals from a U.S. Midwest region—the data source we used for our current analysis is from public profiles posted on the social media website LinkedIn. LinkedIn is a professional networking site with more than 332 million users worldwide.<sup>1</sup> LinkedIn captures information on over 80% of individuals in the U.S. IT labor force; and the correlation between IT employment numbers generated using LinkedIn data and total employment numbers reported by the U.S. Bureau of Labor Statistics for the “packaged software industry (Standard Industrial Classification 7372), in which a very large fraction of employees are IT employees, is 0.81.”<sup>13</sup> LinkedIn is thus a very comprehensive source for information about IT workers.

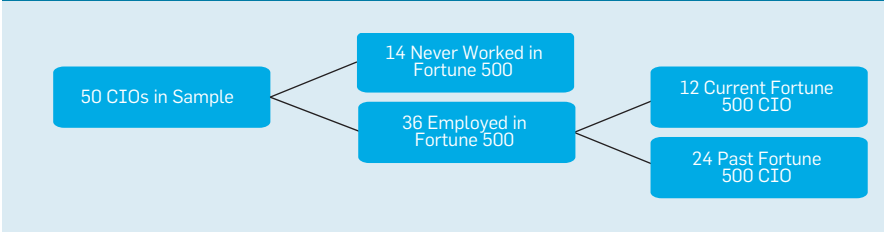
As in the Tanniru study in the 1980s, the detailed résumés of current IT pro-

professionals provided the raw data (such as professional backgrounds and positions held at various points in their careers). The methodology used then and in our current study was to examine the career paths pursued by the IT professionals to inductively derive defining characteristics. The participants in our study reported their professional information on profiles posted in LinkedIn, including employment histories, education, geographic locations, accomplishments, and interest groups. The information on the website is provided voluntarily, with each professional choosing to provide the information he or she deems appropriate. Different levels of detail are thus provided for each individual.

When evaluating the validity of this data, an important measurement concern for us was the likelihood that IT

professionals report their technical skills accurately. For instance, there may be a tendency among younger IT workers to report online platform skills even if they lack a useful level of proficiency. Similarly, older IT workers, with extensive backgrounds in IT, may post only a few of their technical skills. There is also some concern about the possibility of fake profiles on social networks.<sup>16</sup> However, such outright lying and other misrepresentations can have serious repercussions when discovered by someone inside or outside the firm.<sup>8</sup> Additional influences that keep résumé embellishment at bay include the implicit checks associated with peer monitoring and the potential for public embarrassment if one is caught lying in highly scrutinized public social profiles. In fact, a 2012 study by Guillory and Hancock<sup>7</sup> concluded

**Figure 1. CIO sample summary.**



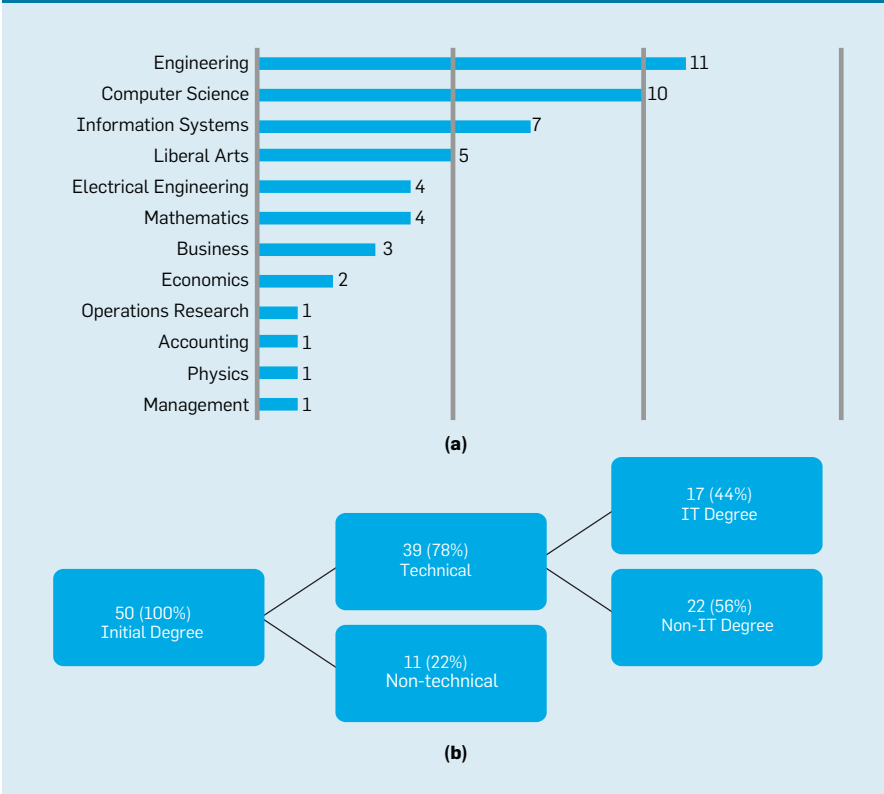
**Figure 2. Enumeration of degrees and majors.**

Degrees	Count	% Total	Majors	Count	% Total	Mappings		
						Tech	NonTech	IT Major
Associates	1	1%	Computer Science	15	16%	X		X
Bachelor of Arts	6	7%	Information Systems	15	16%	X		X
Bachelor of Science	45	49%	Accounting	1	1%	X		
Master of Healthcare Admin.	2	2%	Electrical Engineering	4	4%	X		
Master of Business Admin.	23	25%	Engineering	16	17%	X		
Master of Science	11	12%	Mathematics	4	4%	X		
Juris Doctor	1	1%	Operations Research	3	3%	X		
Doctor of Philosophy	3	3%	Physics	1	1%	X		
	92	100%	Business	22	24%		X	
			Economics	2	2%		X	
			Healthcare	2	2%		X	
			Law	1	1%		X	
			Liberal Arts	5	5%		X	
			Management	1	1%		X	
				92	100%			

Degrees	Count	% Total
Associates	1	1%
Bachelors	51	55%
Masters	36	39%
Doctoral	4	4%
	92	100%

**Figure 3. First degree earned.**



“... web sites such as LinkedIn, which make résumé information public and linked to one’s network, can foster greater honesty for résumé claims that are most important to employers.”

**Data Gathering**

Searching for CIOs in LinkedIn is not straightforward. The search engine explores only the user’s direct connections and up to three connections away from the user. It does not perform a global search of every user profile on LinkedIn. For this reason, we selected a known list of CIOs from top U.S. corporations, and searched for their public LinkedIn profiles using their names. The names came from *The Wall Street Journal’s* 2014 CIO Network Membership List<sup>17</sup> that included more than 100 well-known chief information and chief technology officers from the world’s largest companies, providing a valid and vetted population for studying CIO career paths. In 2014, *The Wall Street Journal’s* CIO Network also posted a biography of each member, allowing for cross-validation of information in each person’s LinkedIn profile.

LinkedIn had public profiles for 107 of the 137 CIOs on the list. Of these, only 50 were complete enough for us to use in our study. We judged profiles to be unusable for such reasons as missing graduation dates from college, no starting or ending dates for job positions, incomplete job titles or roles, or failing to specify degrees or majors for their degrees. This information is needed to ensure consistency and comparability in the data. When constructing career paths, we excluded jobs considered voluntary or community-service oriented (such as time donated to religious or community-service organizations) from the analysis. In total, we collected 50 CIO profiles with complete data during December 2014. These profiles, in aggregate, reflected 319 different job experiences encompassing 1,269 person-years of work experience. Of the 319 experiences, 124, or almost 40%, were with a Fortune 500 firm.

As shown in Figure 1, not all of the CIOs selected for our current study had worked at a Fortune 500 company. When we collected the data, some were working for Fortune 500 companies, some had previously worked for Fortune 500 companies, and some



had never worked for Fortune 500 companies. But all of them, as of December 2014, held the title of CIO, providing a common starting point for us to seek meaningful insights in this exploratory study.

**Educational Background of CIOs**

To summarize the educational backgrounds of the sampled CIOs, we mapped the raw data into established academic degree and major categories (see Figure 2). All the degrees earned by the 50 CIOs in our sample mapped into one of these eight degree categories and 14 major categories. We further characterized the majors as being either technical or non-technical in nature. The computer science and information systems majors, because they specifically educate their graduates for careers in IT, were characterized as IT majors for the analysis.

The 50 CIOs in the dataset were very well educated, having earned a total of 92 degrees, as in Figure 2. A majority (66%) had earned a master’s and/or doctoral degree. In addition, three of the CIOs earned two bachelor’s degrees, and three earned two master’s degrees.

**Initial Degree of CIOs**

The first degree earned can have a significant influence on future career opportunities. As shown in Figure 3, 78% of the first degrees earned by the CIOs were technical in nature, and only 22% non-technical. Engineering, computer science, and information systems were the most frequently chosen undergraduate majors for CIOs. This leads us to conclude that obtaining a technical degree is an important rung on the CIO ladder. Moreover, of those who earned a technical degree, 44% of the CIOs were trained in IT; that is, earned a bachelor’s degree in either computer science or information systems.

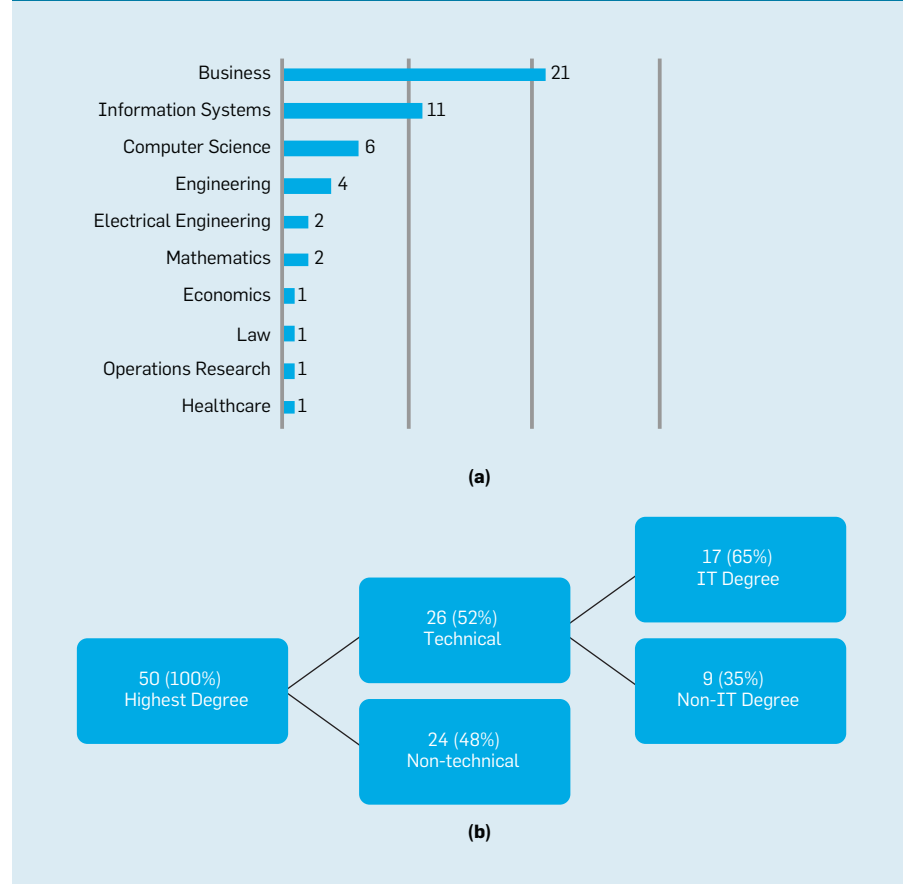
**Highest Degree of CIOs**

Of the 50 CIOs we considered, 66% earned an advanced degree; Figure 4 shows the highest degree earned by each of them. Business administration was by far the most popular major for the last degree earned (42%), marking a shift from its popularity for the first degree earned (6%). The choice of a technology major shifted from 78% for the first degree earned to

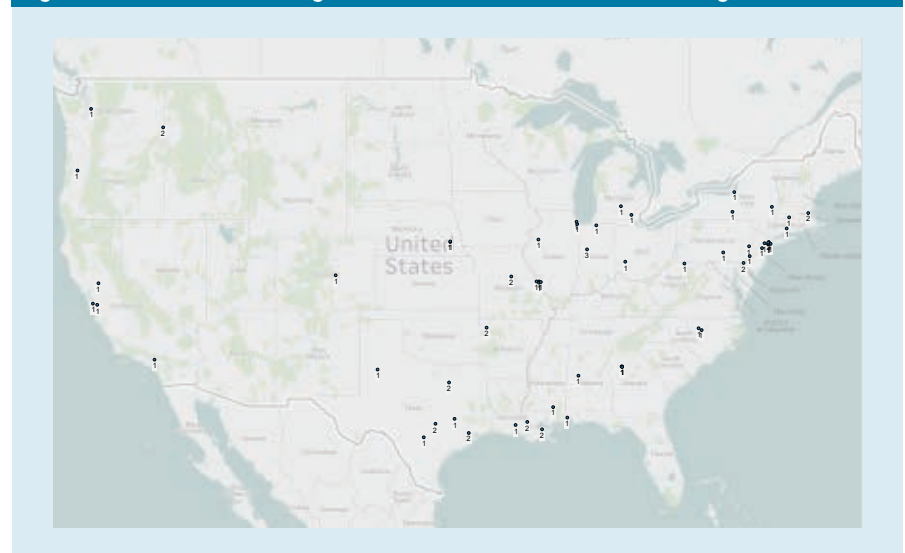
52% for the last degree earned. While the popularity of an IT major stayed at 34% for both the first and last degrees earned, the information systems major was more popular for the last degree than the first (22% vs. 12%), and the computer science major was more popular for the first degree than the last (20% vs. 14%).

From a discipline perspective, engineering and computer science, which were the most popular majors for the first degree earned (42%), were replaced by business administration and information systems for the last degree earned (selected 64% of the time). Of the 29 CIOs whose final degree was a master’s, 70% earned an MBA and

**Figure 4. Highest degree earned.**



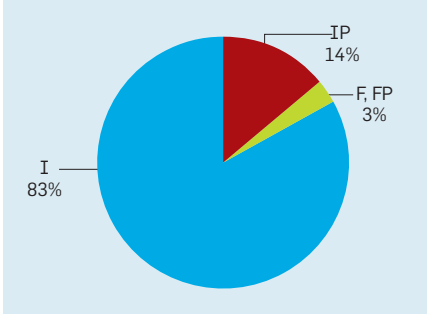
**Figure 5. Location of U.S. colleges and universities where CIOs earned degrees.**



**Figure 6. Service role classification.**

Job Classification		Customer Service Role	
		Internal to the Firm	External to the Firm
Position	IT	Traditional IT (I)	IT Partner (IP)
	Non-IT	Firm (F)	Firm Partner (FP)

**Figure 7. Percentage years of experience by job role.**



28% earned a master's in information systems. Only one of the 16 CIOs with a terminal master's degree and a technical undergraduate degree earned a technical master's degree. This led us to conclude that an advanced degree is an important rung on the CIO ladder, and that an advanced degree should be business oriented.

**Educational Institutions**

An interesting insight we gleaned from the educational background of the CIOs is that achieving a leadership role in IT does not appear to be affected by attending a particular school. The 92 degrees earned by the 50 CIOs we looked at came from 74 different institutions from around the world, including public, private, highly ranked, unranked, well known, and unknown; 64% of the degrees came from U.S. institutions. Interestingly, 66 of the 74 institutions awarded a degree to just one CIO, only seven schools awarded degrees to two CIOs, and just one school (London Imperial College) awarded degrees to three CIOs. Figure 5 shows the locations of schools in the U.S. that awarded degrees to the CIOs in our sample. Note no single school or set of schools in a particular geographic region led the way in educating these CIOs. This led us to conclude that getting a degree from a particular school or region is not an important rung on the CIO ladder.

**Career Transitions**

When the 50 CIOs in our sample began their careers, the top three industries employing them were information technology and services (24%), telecommunications (12%), and defense and space (10%). At the time they became CIOs, the top four industries that employed them were pharmaceuticals (8%), insurance (8%), financial services (8%), and information and technology services (8%). For the CIOs in our sample, the industry in which they became a CIO frequently was not the industry in which they began their careers.

Examining the career paths of IT professionals who achieved the CIO position showed they made career transitions across positions, roles, organizations, and industries. To understand these career paths, it is critical that these positions be mapped to defined roles. IT professionals are considered service providers. The Information Technology Infrastructure Library (ITIL) defines a framework that distinguishes between internal and external IT service providers.<sup>6</sup> With this framework, service providers can be categorized as IT and non-IT and internally and externally focused, as in Figure 6.

In Figure 7, internal IT service providers are labeled as having traditional IT roles, external IT service providers

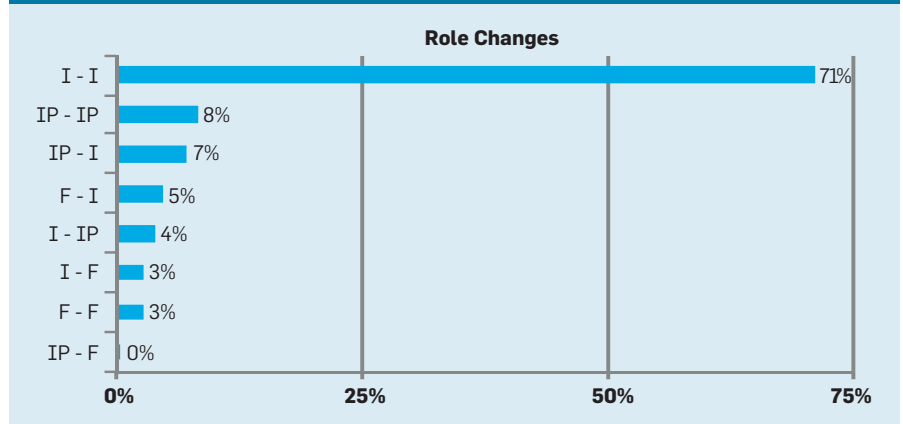
are labeled as having IT partner roles, internal non-IT service providers are labeled as having firm roles, and external non-IT service providers are labeled as having firm partner roles. IT partners function as IT consultants or IT talent for hire to a third party. Firm partners function as non-IT consultants or non-IT talent for hire to a third party. With this classification, it is easy to track movement from an IT position to a non-IT position and firms that supply IT services and firms that buy IT services.

**Transition Across Roles**

Using the roles in Figure 6—"traditional IT (I)," "IT partner (IP)," "firm (F)," and "firm partner (FP)"—the 1,270 person years of experience we collected can be broken down as 83% of the years in traditional IT roles, 14% of the years in IT partner roles, and 3% of the years in non-IT roles (see Figure 8). Breakdown of the data by individual shows only 8% of the CIOs in the sample had experience in non-IT, traditional IT, and IT partner roles (all three roles); approximately 25% of the CIOs had experience in both traditional IT and IT partner roles (both roles); and over 50% of the CIOs spent their entire careers in traditional IT roles (a single role).

Figure 8 tracks changes in roles for each reported job experience. Note that over 70% of job changes resulted in moving from a traditional IT role to another traditional IT role. The next most frequent job changes involved moving from an IT partner role to another IT partner role (8% of moves) and moving from an IT partner role to a traditional IT role (7% of moves). Only 8% of the role changes were associated

**Figure 8. Role changes associated with job changes.**



with moving from a non-IT position to an IT position or from an IT position to a non-IT position. The vast majority of steps up the IT career ladder involve moving from one IT position to another IT position.

**Transition Across Jobs, Organizations, Industries**

We defined a job change as when someone either receives a new job title in the same organization or changes organization. During their careers, the CIOs in our sample changed jobs every four years (47.8 months) on average. The average number of positions held by the IT professionals in the sample before becoming a CIO was 5.1. We concluded that an IT professional with CIO aspirations should not stay in one position for a long period of time but rather should change positions periodically in order to move up the IT career ladder.

When looking for a CIO, organizations may prefer candidates with experience from the same industry, since they bring domain experience associated with that industry. On the other hand, organizations also might prefer candidates from a different industry, since they could bring a new perspective. As of December 2014, LinkedIn identified 147 distinct industries.<sup>10</sup> While this list is not associated with any national or international industry classification scheme (such as the North American Industry Classification System), it does constitute a logical, mutually exclusive and exhaustive grouping. LinkedIn keeps its industry designation consistent by allowing only registered representatives of an organization to choose the industry designation. That is, individuals creating a personal profile do not choose the industry for their organization; the designation is assigned instead by the LinkedIn authorized organization representative.<sup>9</sup>

The 50 CIOs in our sample accumulated 1,269 person-years of work experience in 52 unique industries. From the perspective of total time spent, 17% of the years were in information technology and services, 10% were in financial services, 5.7% were in telecommunications, and 5% were in defense and space. Of the 52 industries, 10 accounted for over 60% of the years of

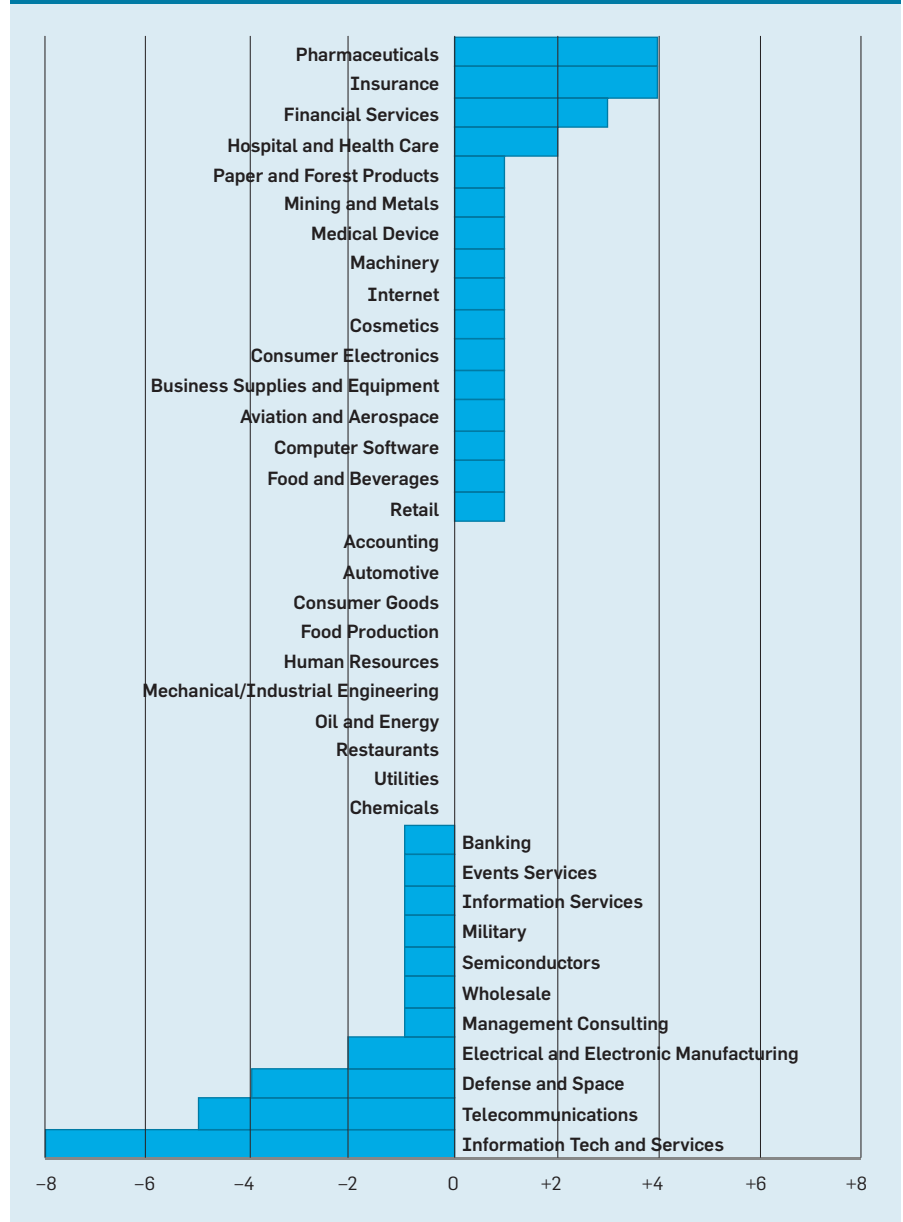
experience of these CIOs. The remaining 42 accounted for 40% of the years of experience in total, with no single industry accounting for more than 2.7% of the years.

On average, the CIOs in the sample made 1.5 industry changes prior to attaining their first CIO position. After attaining that position, the CIOs in the sample made an average of 1.2 industry changes. Because all the CIOs in the sample were still working as of December 2014, and some had just become a CIO, this history indicated that industry changes by IT professionals may not be less frequent after they attain the title of CIO. We also note that 75% of the IT professionals who changed

organizations prior to attaining a CIO title also changed the industry in which they worked. Overall, we found no discernible influence of industry on career progression, implying organizations may be industry agnostic and IT professionals are not tied to any particular industry.

For the 50 CIOs in the sample, the industries in which they began their careers were quite different from the industries in which they were employed in December 2014 when we collected our data. Figure 9 shows that over the course of their careers, these IT professionals tended to move out of information technology and services, telecommunication, defense and

**Figure 9. Net movement between industries during careers.**



space, and electrical and electronic manufacturing and into pharmaceuticals, insurance, financial services, and hospital and healthcare. In Figure 9, the horizontal axis measures the net change in the number of persons in the sample who began their careers in the industry and the number who were in the industry at the time (December 2014) we collected the data for our sample. Although this movement most likely reflects changes in demand for IT services by the various industries rather than preference for a particular industry by IT professionals, it clearly shows that movement between industries is common for IT professionals moving up the career ladder toward a CIO position.

**Time to the Top**

For the 50 CIOs in the sample, it took

an average of 21.0 years from the date their first degree was awarded to the date they attained their first CIO position. Figure 10 shows that the time it took to reach the CIO position is correlated with the decade when the first degree was conferred—1970s, 1980s, or 1990s. Note the steep decline in the amount of time required. A possible explanation for this decline is provided in the following section.

**Stability at the Top**

While IT professionals go through many different positions before they reach their first CIO position, their propensity to change from a CIO position to a non-CIO position appears to be quite low. In our sample, comparing the job experiences of IT professionals before and after they became CIOs showed a 78% decrease in the

average number of role changes per year, a 58% decrease in the average number of organizational changes per year, and a 61% decrease in the average number of industry changes per year. For the 50 CIOs in the sample, Figure 11 shows the distribution of the number of years as a CIO. On average, the IT professionals in the sample stayed in a CIO type role for 8.5 years, which is an underestimate of the true time spent as a CIO since the sample included active CIOs. It appears that once IT professionals take on a CIO position, they generally stay on as CIO (in the same company or in a different company) or retire.

**Promoted from Within vs. Brought In from Outside**

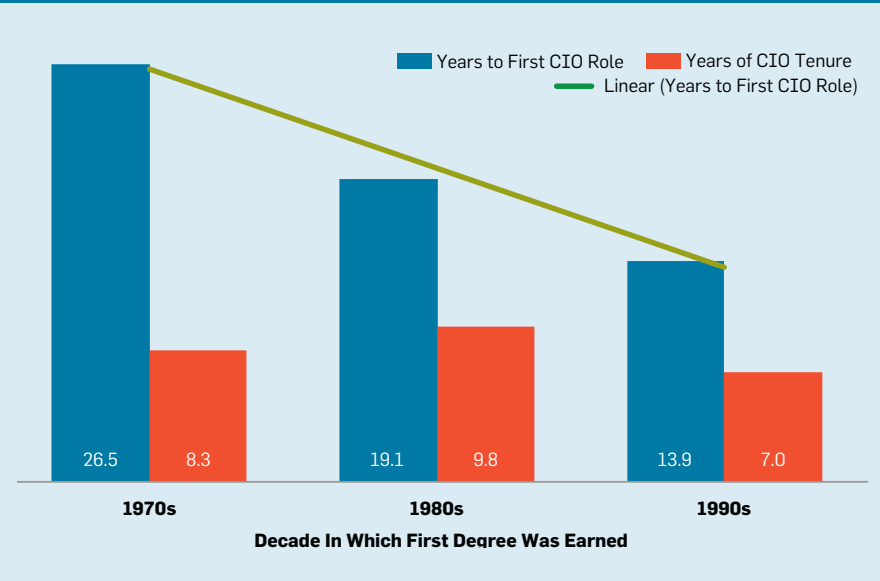
Hiring a CIO from within signals the existence of career paths to the top, while bringing in top talent from outside signals a need for varied experience. The reasons to go outside a firm can include a desire for new methods, new ideas, or even a change in leadership, with the new CEO bringing in a new team. IT Professionals must decide whether the best path to reach a leadership position is to stay within a firm or look elsewhere for the next step in their career progression.

The data in our sample points to IT professionals looking outside their current firms, since 56% changed organizations to obtain their first CIO role and only 44% were promoted from within. Of those promoted from within, all but one were promoted from an internal IT position. The lone exception was a director of finance to being promoted to CIO. However, she had a strong IT background prior to the finance position and had been in finance for only one year.

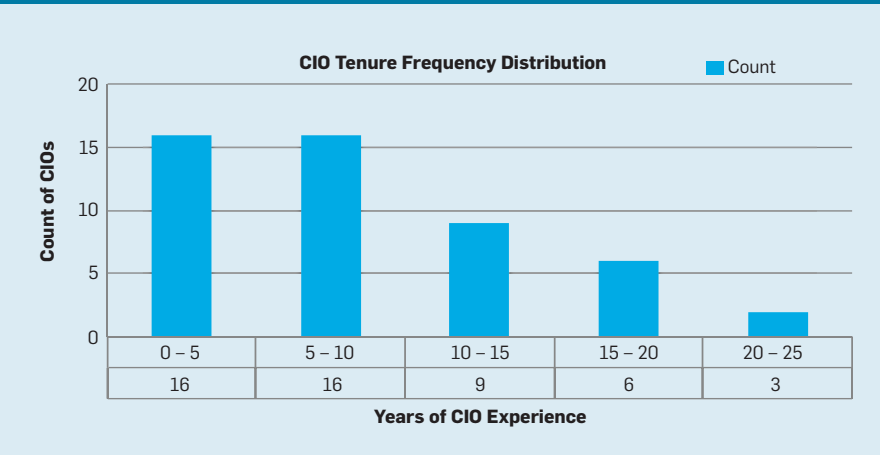
Of those recruited from outside for their first CIO position, we observed two interesting patterns. The first involved role changes. Surprisingly, 30% of the IT professionals who changed organizations for their first CIO opportunity also changed from having an IT partner position to having a traditional IT role. Organizations are apparently comfortable hiring IT professionals from firms with which they partner.

The second involved industry changes. When a CIO is hired, the

**Figure 10. The declining amount of time needed to attain a CIO position.**



**Figure 11. Distribution of years spent as a CIO.**



firm must decide if candidates from the same industry are more attractive (because they have similar domain knowledge) or if IT professionals from other industries are more attractive (because they have a wider range of experience). The data in our study shows 75% of the IT professionals who changed organizations in order to obtain their first CIO position also changed their industry. This indicates that firms are looking for professionals with diverse IT experience or that IT may be industry agnostic. In either case, it appears that IT professionals are not tied to an industry.

### Changing Nature of IT

Information technology is constantly changing. Advances in the digitization of business processes/services, the miniaturization of digital products that store and manipulate data, and the use of agile development methods that deliver applications faster to meet business needs, are just a few examples of such changes. One by-product in IT is an increasing specialization of roles among those who design, develop, and implement IT.

In-depth interviews by Tanniru<sup>15</sup> with more than 80 IT executives in 2012 showed a stark change in the focus of IT leadership over the past four decades—from focusing primarily on costs to focusing on strategic alignment and entrepreneurship. In the 1970s and early 1980s, a few large companies produced almost all the IT products worldwide, with IT leaders using these products to seek cost reductions in business operations. The life cycles for these projects tended to be quite long, and the CIO was responsible for managing all phases of the projects.

In the late 1980s and early 1990s, introduction of PCs and a large number of small product vendors and off-the-shelf software tools empowered business users to seek control of their IT resources for decision support. This led to two parallel approaches. The spiral development approach facilitated multiple smaller projects, while the conventional waterfall approach developed and maintained larger projects. CIOs had to become adept at managing both large and small projects.

In the late 1990s and early 2000s



**In the new environment, talent can be recognized sooner than in the past, and there are more opportunities to display leadership as IT becomes a strategic tool for competitiveness in the marketplace.**



another major change took place. Large integrated enterprise software packages commoditized many business-process support tools, empowering business and IT managers to outsource these tools to external vendors/partners for improved productivity and customer responsiveness. More often than not, the middle part of the development life cycle—design and develop—was outsourced to external enterprise software vendors. CIOs had to become adept at managing large and small projects from multiple vendors, by focusing on planning, analysis, implementation, and integration.<sup>15</sup>

As we entered the 21<sup>st</sup> century, the focus of IT management shifted to developing newer applications using advanced Web- and social media-based technologies. The advances made in Internet and Web-based technologies in the 2000s led to a plethora of hardware, software, and service companies, all serving various parts of a firm's extended value chain and empowering customers to seek improved services. As a result, CIOs today must manage two different life cycles: one looks at digital innovations to explore newer technologies, while the other continues to support legacy systems. Multiple partners are used to assist with both these life cycles. Today's CIOs must be adept at keeping the lights on and quick pilot testing of new digital innovations. They have to balance the strategic needs of a changing business through exploration while maintaining a reliable backend operational system using a mix of both established and innovative technology vendors/partners, or two-speed IT.<sup>2</sup>

These changes in IT led to the establishment of two distinct vendor groups, in addition to the traditional IT user within a firm. One group of vendors continues to advance IT innovation with new products and services (such as social media, Internet of Things, and data analytics) in support of changing business needs, while the other continues to implement and maintain commoditized IT products or services (such as enterprise systems, data warehouses, Web services, and standardized hardware plug-ins). The IT leadership within a firm is expected to interact with both types of vendors

while ensuring internal IT continues to support the various applications that are critical to the business.

An interesting by-product of this evolving IT development cycle is the opportunities it creates for CIOs to display leadership. The shorter cycle times associated with the highly visible projects that support innovation allow IT professionals to gain more experience in integrating business and IT, and become more visible to senior leadership in a much shorter period of time. Increased visibility is a contributing factor in the reduction of the amount of time it takes an IT professional to attain the position of CIO.

### Conclusion

Employee retention and satisfaction are closely tied to career path development and progression.<sup>18</sup> We examined the paths several prominent IT professionals followed to achieve a CIO position. This information should be of interest to those entering the IT field with aspirations to become a CIO. The information should also be of interest to IT leaders recruiting new IT talent who want to help that talent develop. However, the reader is cautioned that the findings from this study are for a select group of CIOs and may not be generalizable to all CIOs.

Several patterns emerged from our look at specific career paths. From an education perspective, we observed that the majority of CIOs in our sample earned a bachelor's degree in a technical field and a master's degree in a business field. As in Bruni,<sup>3</sup> our results show the school at which the degree is earned does not appear to have any influence on becoming a CIO. However, it is important to have both a bachelor's degree and a master's degree, and it is just as important to have both technical and business knowledge.

Another interesting pattern is that experience in a particular industry does not seem to affect whether one becomes a CIO. The fundamental IT skills appear to be applicable to any industry. On the path to the top, CIOs often changed both industry and organization. In fact, of those who changed organizations to attain their first CIO position, 75% also changed industries. This implies that either IT skills are industry agnostic, or that tacit industry

knowledge is not a significant factor leading to CIO success.

The path to the top for the CIOs in our sample was also characterized by frequent position and organization changes. Consistent with conventional wisdom, moving to new positions and firms appears to help mold future leaders. However, contrary to conventional wisdom,<sup>12</sup> our data shows that one does not need to have line-of-business experience to attain the CIO position, or even experience as an IT consultant. The majority of the years of experience of the CIOs in our sample was spent in traditional IT positions (83%), while only 14% of the years was in consulting positions, and only 3% of the years was in non-IT positions. Many CIOs do not have any experience outside of IT or in organizations that provide IT services. This indicates that neither non-IT functional area roles nor IT service provider roles are prerequisites for a CIO position.

Finally, we observed that modern IT structures and rapid change have shortened the time it takes to become a CIO by nearly 50% compared to 30 years ago. In the new environment, talent can be recognized sooner than in the past, and there are more opportunities to display leadership as IT becomes a strategic tool for competitiveness in the marketplace. That is, modern IT strategies provide opportunities to demonstrate significant leadership ability earlier in one's career, resulting in a much more rapid ascent to the CIO position.

These initial results can be neatly summarized into "needs" and "need nots" for aspiring CIOs. They definitely need technical expertise, a master's degree, and to change positions and companies. However, they need not go to a particular school, stay with the same company, stay in the same industry, or gain non-IT experience. If IT professionals follow these needs and need nots, they can become a CIO in a surprisingly short period of time.

Our exploratory research produced interesting and useful information about the characteristics of a specific group of successful CIOs, including their career paths, educational backgrounds, and cross-industry and intra-firm experiences. Yet more research is warranted. A primary objective would be to study a large, random sample of

CIOs from, say, the Fortune 500, Fortune 1,000, governmental, educational, nonprofit, and small and medium-size business sectors. This would enable us to move from descriptive statistics to predictive and prescriptive analyses. □

### References

1. Blake, K.E. *The 2015 LinkedIn Statistics That You Need to Know* (Jan. 29, 2015); <https://www.linkedin.com/pulse/2015-linked-in-statistics-you-need-know-katy-elle-blake>
2. Bossert, O., Ip, C., and Laartz, J. A two-speed IT architecture for the digital enterprise. McKinsey & Company, Dec. 2014; [http://www.mckinsey.com/insights/business\\_technology/a\\_two\\_speed\\_it\\_architecture\\_for\\_the\\_digital\\_enterprise](http://www.mckinsey.com/insights/business_technology/a_two_speed_it_architecture_for_the_digital_enterprise)
3. Bruni, F. *Where You Go Is Not Who You'll Be: An Antidote to the College Admissions Mania*. Hachette Book Group, New York, 2015.
4. Brynjolfsson, E., Hitt, L., and Yang, S. Intangible assets: Computers and organizational capital. *Brookings Papers on Economic Activity* 2002, 1 (Jan. 2002), 137–181.
5. Desmet, D., Duncan, E., Scanlan, J., and Singer, M. Six building blocks for creating high-performing digital enterprises. McKinsey & Company, Sept. 2015; <http://www.mckinsey.com/business-functions/organization/our-insights/six-building-blocks-for-creating-a-high-performing-digital-enterprise>
6. Great Britain Office of Government Commerce. *Service Strategy*. The Stationery Office, London, U.K., 2007.
7. Guillory, J. and Hancock, J.T. The effect of LinkedIn on deception in résumés. *Cyberpsychology, Behavior, and Social Networking* 15, 3 (Mar. 2012), 135–140.
8. Kidwell Jr., R.E. 'Small' lies, big trouble: The unfortunate consequences of résumé padding, from Janet Cooke to George O'Leary. *Journal of Business Ethics* 51, 2 (May 2004), 175–184.
9. LinkedIn. Company pages; <http://press.linkedin.com/about-linkedin>
10. LinkedIn. Industry codes; <http://developer.linkedin.com/docs/reference/industry-codes>
11. Melville, N., Kraemer, K., and Gurbuxani, V. Review: Information technology and organizational performance: An integrative model of IT business value. *MIS Quarterly* 28, 2 (June 2014), 283–322.
12. Smith, G.S. *Straight to the Top: CIO Leadership in a Mobile, Social, and Cloud-based World, 2nd Edition*. John Wiley & Sons, New York, 2013, 110–113.
13. Tambe, P. Big data investment, skills, and firm value. *Management Science* 60, 6 (June 2014), 1452–1469.
14. Tanniru, M.R. An investigation of the career path of the EDP professional. In *Proceedings of the 20th Annual Computer Personnel on Research Conference* (Charlottesville, VA, Nov. 17–18). ACM Press, New York, 1983, 87–101.
15. Tanniru, M.R. Should IS departments have a strong presence in the business school? *ACM SIGMIS Database* 43, 2 (May 2012), 15–19.
16. Thier, D. An estimated 83 million Facebook profiles are fake. *Forbes* (Aug. 2, 2012); <http://www.forbes.com/sites/davidthier/2012/08/02/83-million-estimated-facebook-profiles-are-fake/>
17. *The Wall Street Journal*. CIO Network (Nov. 11, 2014); <http://cionetwork.wsj.com>
18. Woods, K. Exploring the relationship between employee turnover rate and customer satisfaction levels. *The Exchange* 4, 1 (Sept. 2015), 33–43.

**Daniel J. Mazzola** (dan.mazzola@asu.edu) is a clinical assistant professor in the Information Systems Department of the W. P. Carey School of Business at Arizona State University, Tempe, AZ.

**Robert D. St. Louis** (st.louis@asu.edu) is a professor in the Information Systems Department of the W. P. Carey School of Business at Arizona State University, Tempe, AZ.

**Mohan R. Tanniru** (tanniru@oakland.edu) is a professor of MIS in the Decision and Information Science Department of the School of Business Administration at Oakland University, Rochester, MI.



## New tools tackle an age-old practice.

BY SIMON PRICE AND PETER A. FLACH

# Computational Support for Academic Peer Review: A Perspective from Artificial Intelligence

PEER REVIEW IS the process by which experts in some discipline comment on the quality of the works of others in that discipline. Peer review of written works is firmly embedded in current academic research practice where it is positioned as the gateway process and quality control mechanism for submissions to conferences, journals, and funding bodies across a wide range of disciplines. It is probably safe to assume that peer review in some form will remain a cornerstone of academic practice for years to come, evidence-based criticisms of this process in computer science<sup>22,32,45</sup> and other disciplines<sup>23,28</sup> notwithstanding.

While parts of the academic peer review process have been streamlined in the last few decades to take technological advances into account, there are many

more opportunities for computational support that are not currently being exploited. The aim of this article is to identify such opportunities and describe a few early solutions for automating key stages in the established academic peer review process. When developing these solutions we have found it useful to build on our background in machine learning and artificial intelligence: in particular, we utilize a feature-based perspective in which the handcrafted features on which conventional peer review usually depends (for example, keywords) can be improved by feature weighting, selection, and construction (see Flach<sup>17</sup> for a broader perspective on the role and importance of features in machine learning).

Twenty-five years ago, at the start of our academic careers, submitting a paper to a conference was a fairly involved and time-consuming process that roughly went as follows: Once an author had produced the manuscript (in the original sense, that is, manually produced on a typewriter, possibly by someone from the university's pool of typists), he or she would make up to seven photocopies, stick all of them

### » key insights

- **State-of-the-art tools from machine learning and artificial intelligence are making inroads to automate parts of the peer-review process; however, many opportunities for further improvement remain.**
- **Profiling, matching, and open-world expert finding are key tasks that can be addressed using feature-based representations commonly used in machine learning.**
- **Such streamlining tools also offer perspectives on how the peer-review process might be improved: in particular, the idea of profiling naturally leads to a view of peer review being aimed at finding the best publication venue (if any) for a submitted paper.**
- **Creating a more global embedding for the peer-review process that transcends individual conferences or conference series by means of persistent reviewer and author profiles is key, in our opinion, to a more robust and less arbitrary peer-review process.**





in a large envelope, and send them to the program chair of the conference, taking into account that international mail would take 3–5 days to arrive. On their end, the program chair would receive all those envelopes, allocate the papers to the various members of the program committee, and send them out for review by mail in another batch of big envelopes. Reviews would be completed by hand on paper and mailed back or brought to the program committee meeting. Finally, notifications and reviews would be sent back by the program chair to the authors by mail. Submissions to journals would follow a very similar process.

It is clear that we have moved on quite substantially from this paper-based process—indeed, many of the steps we describe here would seem arcane to our younger readers. These days, papers and reviews are submitted online in some conference management system (CMS), and all communication is done via email or via message boards on the CMS with all metadata concerning people and papers stored in a database backend. One could argue this has made the process much more efficient, to the extent that we now specify the submission deadline up to the second in a particular time zone (rather than approximately as the last post round at the program chair’s institution), and can send out hundreds if not thousands of notifications at the touch of a button.

Computer scientists have been studying automated computational support for conference paper assignment since pioneering work in the 1990s.<sup>14</sup> A range of methods have been used to reduce the human effort involved in paper allocation, typically with the aim of producing assignments that are similar to the ‘gold standard’ manual process.<sup>9,13,16,18,30,34,37</sup> Yet, despite many publications on this topic over the intervening years, research results in paper assignment have made relatively few inroads into mainstream CMS tools and everyday peer review practice. Hence, what we have achieved over the last 25 years or so appears to be a streamlined process rather than a fundamentally improved one: we believe it would be difficult to argue the decisions taken by program committees today are significantly better in comparison with the paper-based process. But this doesn’t mean that opportunities for improving the process don’t exist—on the contrary, there is, as we demonstrate in this article, considerable scope for employing the very techniques that researchers in machine learning and artificial intelligence have been developing over the years.

The accompanying table recalls the main steps in the peer review process and highlights current and future opportunities for improving it through advanced computational support. In discussing these topics, it will be helpful to draw a distinction between closed-world and open-world settings. In a closed-

world setting there is a fixed or predetermined pool of people or resources. For example, assigning papers for review in a closed-world setting assumes a program committee or editorial board has already been assembled, and hence the main task is one of matching papers to potential reviewers. In contrast, in an open-world setting the task becomes one of finding suitable experts. Similarly, in a closed-world setting an author has already decided which conference or journal to send their paper to, whereas in an open-world setting one could imagine a recommender system that suggests possible publication venues. The distinction between closed and open worlds is gradual rather than absolute: indeed, the availability of a global database of potential publication venues or reviewers with associated metadata would render the distinction one of scale rather than substance. Nevertheless, it is probably fair to say that, in the absence of such global resources, current opportunities tend to be focus on closed-world settings. Here, we review research on steps II, III and V, starting with the latter two, which are more of a closed-world nature.

### Assigning Papers for Review

In the currently established academic process, peer review of written works depends on appropriate assignment to several expert peers for their review. Identifying the most appropriate set of reviewers for a given submitted paper is a time-consuming and non-trivial task for conference chairs and journal editors—not to mention funding program managers, who rely on peer review for funding decisions. Here, we break the review assignment problem down into its matching and constraint satisfaction constituents, and discuss possibilities for computational support.

Formally, given a set  $P$  of papers with  $|P| = p$  and a set  $R$  of reviewers with  $|R| = r$ , the goal of paper assignment is to find a binary matrix  $A^{r \times p}$  such that  $A_{ij} = 1$  indicates the  $i$ -th reviewer has been assigned the  $j$ -th paper, and  $A_{ij} = 0$  otherwise. The assignment matrix should satisfy various constraints, the most typical of which are: each paper is reviewed by at least  $c$  reviewers (typically,  $c = 3$ ); each reviewer is assigned no more than  $m$  papers, where  $m = O(p/r)$ ; and reviewers should not be assigned

**A chronological summary of the main activities in peer review, with opportunities for improving the process through computational support.**

	Actor	Activity	What can be done now	What might be done in future
I	Author	Paper submission		Recommender systems for publication venue; papers carry full previous reviewing history
II	Program chair	Assembling program committee	Expert finding	PCs for an area rather than a single conference; workload balancing
III	Program chair	Assigning papers for review	Bidding and assignment support	Extending PCs based on submitted papers
IV	Reviewer	Reviewing papers		Advanced reviewing tools that find related work and map the paper under review relative to it
V	Program chair	Discussion and decisions	Reviewer score calibration	More outcome categories; recommender systems for outcomes; more decision time points

papers for which they have a conflict of interest (this can be represented by a separate binary conflict matrix  $C^{r,p}$ ). As this problem is underspecified, we will assume that further information is available in the form of a score matrix  $M^{r,p}$  expressing for each paper-reviewer pair how well they are matched by means of a non-negative number (higher means a better match). The best allocation is then the one that maximizes the element-wise matrix product  $\sum_{ij} A_{ij} M_{ij}$  while satisfying all constraints.<sup>44</sup>

This one-dimensional definition of ‘best’ does not guarantee the best set of reviewers if a paper covers multiple topics, for example, a paper on machine learning and optimization could be assigned three reviewers who are machine learning experts but none who are optimization experts. This shortcoming can be addressed by replacing  $R$  with the set  $R^c$  such that each  $c$ -tuple  $\in R^c$  represents a possible assignment of  $c$  reviewers.<sup>24,25,42</sup> Recent works add explicit constraints on topic coverage to incorporate multiple dimensions into the definition of best allocation.<sup>26,31,40</sup> Other types of constraints have also been considered, including geographical distribution and fairness of assignments, as have alternative constraint solver algorithms.<sup>3,19,20,43</sup> The score matrix can come from different sources, possibly a combination. Here, we review three possible sources: feature-based matching, profile-based matching, and bidding.

**Feature-based matching.** To aid assigning submitted papers to reviewers a short list of subject keywords is often required by mainstream CMS tools as part of the submission process, either from a controlled vocabulary, such as the ACM Computing Classification System (CCS),<sup>3</sup> or as a free-text “folksonomy.” As well as collecting keywords for the submitted papers, taking the further step of also requesting subject keywords from the body of potential reviewers enables CMS tools to make a straightforward match between the papers and the reviewers based on a count of the number of keywords they have in common. For each paper the reviewers can then be ranked in order of the number of matching keywords.

If the number of keywords associated with each paper and each reviewer is not fixed then the comparison may be normalized by the CMS to avoid overly favoring longer lists of keywords. If the overall vocabulary from which keywords are chosen is small then the concepts they represent will necessarily be broad and likely to result in more matches. Conversely, if the vocabulary is large, as in the case of free-text or the ACM CCS, then concepts represented will be finer grained but the number of matches is more likely to be small or even non-existent. Also, manually assigning keywords to define the subject of written material is inherently subjective. In the medical domain, where taxonomic classification schemes are commonplace, it has been demonstrated that different experts, or even the same expert over time, may be inconsistent in their choice of keywords.<sup>6,7</sup>

When a pair of keywords does not literally match, despite having been chosen to refer to the same underlying concept, one technique often used to improve matching is to also match their synonyms or syntactic variants—as defined in a thesaurus or dictionary of abbreviations, for example, treating ‘code inspection’ and ‘walkthrough’ as equivalent; likewise for ‘SVM’ and ‘support vector machine’ or ‘ $\lambda$ -calculus’ and ‘lambda calculus.’ However, if such simple equivalence classes are not sufficient to capture important differences between subjects—for example, if the difference between ‘code inspection’ and ‘walk-through’ is significant—then an alternative technique is to exploit the hierarchical structure of a concept taxonomy in order to represent the distance between concepts. In this setting, a match can be based on the common ancestors of concepts—either counting the number of shared ancestors or computing some edge traversal distance between a pair of concepts, for example, the former ACM CCS concept ‘D.1.6 Logic Programming’ has ancestors ‘D.1 Programming Techniques’ and ‘D. Software,’ both of which are shared by the concept ‘D.1.5 Object-oriented Programming,’ meaning that D.1.5 and D.1.6 have a non-zero similarity because they have common ancestors.

Obtaining a useful representation of concept similarity from a taxonomy is challenging because the measures

tend to assume uniform coverage of the concept space such that the hierarchy is a balanced tree. The approach is further complicated as it is common for certain concepts to appear at multiple places in a hierarchy, that is, taxonomies may be graphs rather than just trees, and consequently there may be multiple paths between a pair of concepts. The situation grows worse still if different taxonomies are used to describe the subject of written works from different sources because a mapping between the taxonomies is required. Thus, it is not surprising that one of the most common findings in the literature on ontology engineering is that ontologies, including taxonomies, thesauri, and dictionaries, are difficult to develop, maintain, and use.<sup>12</sup>

So, even with good CMS support, keyword-based matching still requires manual effort and subjective decisions from authors, reviewers and, sometimes, ontology engineers. One useful aspect of feature-based matching using keywords is that it allows us to turn a heterogeneous matching problem (papers against reviewers) into a homogeneous one (paper keywords against reviewer keywords). Such keywords are thus a simple example of profiles that are used to describe relevant entities (papers and reviewers). Next, we take the idea of profile-based matching a step further by employing a more general notion of profile that incorporates nonfeature-based representations such as bags of words.

**Automatic feature construction with profile-based matching.** The main idea of profile-based matching is to automatically build representations of semantically relevant aspects of both papers and reviewers in order to facilitate construction of a score matrix. An obvious choice of such a representation for papers is as a weighted bag-of-words (see “The Vector Space Model” sidebar). We then need to build similar profiles of reviewers. For this purpose we can represent a reviewer by the collection of all their authored or co-authored papers, as indexed by some online repository such as DBLP<sup>29</sup> or Google Scholar. This collection can be turned into a profile in several ways, including: build the profile from a single document or Web page containing the bibliographic details of the reviewer’s publications (see

a <http://www.acm.org/about/class/> (The examples in this article refer to ACM’s 1998 CCS, which was recently updated.)

# The Vector Space Model

The canonical task in information retrieval is, given a query in the form of a list of words (terms), to rank a set of text documents  $D$  in order of their similarity to the query. In the vector space model, each document  $d \in D$  is represented as the multiset of terms (bag-of-words) occurring in that document. The set of distinct terms in  $D$ , vocabulary  $V$ , defines a vector space with dimensionality  $|V|$  and thus each document  $d$  is represented as a vector  $\vec{d}$  in this space. The query  $q$  can also be represented as a vector  $\vec{q}$  in this space, assuming it shares vocabulary  $V$ . The query and a document are considered similar if the angle  $q$  between their vectors is small. The angle can be conveniently captured by its cosine  $\vec{q} \cdot \vec{d} / (|\vec{q}| \cdot |\vec{d}|)$ , giving rise to the cosine similarity.

However, if raw term counts are used in vectors  $\vec{q}$  and  $\vec{d}$  then similarity will: (i) be biased in favor of long documents and; (ii) treat all terms as equally important, irrespective of how commonly they occur across all documents. The *term frequency-inverse document frequency* (tf-idf) weighting scheme compensates for (i) by normalizing term counts within a document by the total number of terms in that document, and (ii) by penalizing terms that occur in many documents, as follows. The *term frequency* of term  $t_i$  in the document  $d_j$  is  $tf_{ij} = n_{ij} / \sum_k n_{kj}$ . The *inverse document frequency* of term  $t_i$  is  $idf_i = \log(|D| / df_i)$ , where *term count*  $n_{ij}$  is the number of times term  $t_i$  occurs in the document  $d_j$ , and *document frequency*  $df_i$  of term  $t_i$  is the number of documents in  $D$  in which term  $t_i$  occurs. A term that occurs often in a document has high term frequency; if it occurs rarely in other documents it has high inverse document frequency. The product of the two, tf-idf, thus expresses the extent to which a term characterizes a document relative to other documents in  $D$ .

# Toronto Paper Matching System

The Toronto Paper Matching System TPMS (papermatching.cs.toronto.edu) originated as a standalone paper assignment recommender for the NIPS 2010 conference and was subsequently loosely integrated with Microsoft's Conference Management Toolkit (CMT) to streamline access to paper submissions for ICML 2012. TPMS requires reviewers to upload a selection of their own papers, reports and other self-selected textual documents, which are then analyzed to produce their reviewer profile. This places control over the scope of the profile in the hands of the reviewers themselves so that they need only include publications about topics they are prepared to review. Once uploaded, TPMS persists the documents and resultant profile beyond the scope of a single conference, allowing reviewers to reuse the same profile for future conferences, curating their own set of characteristic documents as they see fit.

The scoring model used is similar to the vector-space model but takes a Bayesian approach. In addition, profiles in TPMS can be expressed over a set of hypothesized topics rather than raw terms. Topics are modeled as hidden variables that can be estimated using techniques such as Latent Dirichlet Allocation.<sup>4,8</sup> This increased expressivity comes at the cost of requiring more training data to stave off the danger of overfitting.

“SubSift and MLj-Matcher” sidebar); or retrieve or let the reviewer upload full-text of (selected) papers, which are then individually converted into the required representation and collectively averaged to form the profile (see “Toronto Paper Matching System” (TPMS) sidebar). Once both the papers and the reviewers have been profiled, the score matrix  $M$  can be populated with the cosine similarity between the term weight vectors of each paper-reviewer pair.

Profile-based methods for matching papers with reviewers exploit the intuitive idea that the published works of reviewers, in some sense, describe their specific research interests and expertise. By analyzing these pub-

lished works in relation to the body as a whole, discriminating profiles may be produced that effectively characterize reviewer expertise from the content of existing heterogeneous documents ranging from traditional academic papers to websites, blog posts, and social media. Such profiles have applications in their own right but can also be used to compare one body of documents to another, ranking arbitrary combinations of documents and, by proxy, individuals by their similarity to each other.

From a machine learning point of view, profile-based matching differs from feature-based matching in that the profiles are constructed in a data-driven way without the need to come up with a

set of keywords. However, the number of possible terms in a profile can be huge and so systems like TPMS use automatic topic extraction as a form of dimensionality reduction, resulting in profiles with terms chosen from a limited number keywords (topics). As a useful by-product of profiling, each paper and each reviewer is characterized by a ranked list of terms which can be seen as automatically constructed features that could be further exploited, for instance to allocate accepted papers to sessions or to make clear the relative contribution of individual terms to a similarity score (see “SubSift and MLj Matcher” sidebar).

**Bidding.** A relatively recent trend is to transfer some of the paper allocation task downstream to the reviewers themselves, giving them access to the full range of submitted papers and asking them to bid on papers they would like to review. Existing CMS tools offer support for various bidding schemes, including: allocation of a fixed number of ‘points’ across an arbitrary number of papers, selection of top  $k$  papers, rating willingness to review papers according to strength of bid, as well as combinations of these. Hence, bidding can be seen as an alternative way to come up with a score matrix that is required for the paper allocation process. There is also the opportunity to register conflicts of interests, if a reviewer's relations with the authors of a particular paper are such that the reviewer is not a suitable reviewer for that paper.

While it is in a reviewer's self-interest to bid, invariably not all reviewers will do so, in which case the papers they are allocated for review may well not be a good match for their expertise and interests. This can be irritating for the reviewer but is particularly frustrating for the authors of the papers concerned. The absence of bids from some reviewers can also reduce the fairness of allocation algorithms in CMS tools.<sup>19</sup> Default options in the bidding process are unable to alleviate this: if the default is “I cannot review this” the reviewer is effectively excluded from the allocation process, while if the default is to indicate some minimal willingness to review a paper the reviewer is effectively used as a wildcard and will receive those papers that are most difficult to allocate.

A hybrid of profile-based matching and manual bidding was explored for

the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining in 2009. At bidding time the reviewers were presented with initial bids obtained by matching reviewer publication records on DBLP with paper abstracts (see “Experience from SIGKDD’09” sidebar for details) as a starting point. Several PC members reported they considered these bids good enough to relieve them from the temptation to change them, although we feel there is considerable scope to improve both the quality of recommendations and of the user interface in future work. ICML 2012 further explored the use of a hybrid model and a pre-ranked list of suggested bids.<sup>b</sup> The TPMS software used at ICML 2012 offers other scoring models for combining bids with profile-based expertise assessment.<sup>8,9</sup> Effective automatic bid initialization would address the aforementioned problem caused by non-bidding reviewers.

### Reviewer Score Calibration

Assuming a high-quality paper assignment has been achieved by means of one of the methods described earlier, reviewers are now asked to honestly assess the quality and novelty of a paper and its suitability for the chosen venue (conference or journal). There are different ways in which this assessment can be expressed: from a simple yes/no answer to the question: “If it was entirely up to you, would you accept this paper?” via a graded answer on a more common five- or seven-point scale (for example, Strong Accept (3); Accept (2); Weak Accept (1); Neutral (0); Weak Reject (−1); Reject (−2); Strong Reject (−3)), to graded answers to a set of questions aiming to characterize different aspects of the paper such as novelty, impact, technical quality, and so on.

Such answers require careful interpretation for at least two reasons. The first is that reviewers, and even area chairs, do not have complete information about the full set of submitted papers. This matters in a situation where the total number of papers that can be accepted is limited, as in most conferences (it is less of an issue for journals). The main reason why raw reviewer scores are problematic is that different reviewers tend to use the scale(s) involved in different ways. For example,

some reviewers tend to stay to the center of the scale while others tend to go more for the extremes. In this case it would be advisable to normalize the scores, for example, by replacing them with  $z$ -

scores. This corrects for differences in both mean scores and standard deviations among reviewers and is a simple example of reviewer score calibration.

In order to estimate a reviewer’s

## SubSift and MLJ-Matcher

SubSift, short for ‘submission sifting’, was originally developed to support paper assignment at SIGKDD’09 and subsequently generalized into a family of Web services and re-usable Web tools ([www.subsift.com](http://www.subsift.com)). The submission sifting tool composes several SubSift Web services into a workflow driven by a wizard-like user interface that takes the Program Chair through a series of Web forms of a paper-reviewer profiling and matching process.

On the first form, a list of PC member names is entered. SubSift looks up these names on DBLP and suggests author pages which, after any required disambiguation, are used as documents to profile the PC members. Behind the scenes, beginning from a list of bookmarks (urls), SubSift’s harvester robot fetches one or more DBLP pages per author, extracts all publication titles from each page and aggregates them into a single document per author. In the next form, the conference paper abstracts are uploaded as a CSV file and their text is used to profile the papers. After matching PC member profiles against paper profiles, SubSift produces reports with ranked lists of papers per reviewer, and ranked lists of reviewers per paper. Optionally, by manually specifying threshold similarity scores or by specifying absolute quantities, a CSV file can be downloaded with initial bid assignments for upload into a CMS.

For the editor-in-chief of a journal, the task of assigning a paper to a member of the editorial board for their review can be viewed as a special case of the conference paper assignment problem (without bidding), where the emphasis is on finding the best match for one or a few papers. We built an alternative user interface to SubSift that supports paper assignment for journals. Known as *MLJ-Matcher* in its original incarnation, this tool has been used since 2010 to support paper assignment for the *Machine Learning* journal as well as other journals.

## Experience from SIGKDD’09

Our own experience with bespoke tools to support the research paper review process started when Flach was appointed, with Mohammed Zaki from Rensselaer Polytechnic Institute, program co-chair of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2009 (SIGKDD’09). The initial SubSift tools were written by members of the Bristol Intelligent Systems Laboratory with external collaborators at Microsoft Research Cambridge. As reported in Flach et al.,<sup>18</sup> the SubSift tools assisted in the allocation of 537 submitted research papers to 199 reviewers.

Using these tools, each reviewer’s bids were initialized using a weighted sum of cosine similarity between the paper’s abstract and the reviewer’s publication titles as listed in the DBLP computer science online bibliography,<sup>30</sup> and the number of shared subject areas (keywords). The combined similarity scores were discretized into four bins using manually chosen thresholds, with the first bin being a 0 (no-bid) and the other three being bids of increasing strength: 1 (at a pinch), 2 (willing) and 3 (eager). These initial bids were exported from SubSift and imported into the conference management tool (Microsoft CMT, [cmt.research.microsoft.com](http://cmt.research.microsoft.com)).

Based on the same similarity information, each reviewer was sent an email containing a link to a personalized SubSift generated Web page listing details of all 537 papers ordered by initial bid allocation or by either of its two components: keyword matches or similarity to their own published works. The page also listed the keywords extracted from the reviewer’s own publications and those from each of the submitted papers. Guided by this personalized perspective, plus the usual titles and abstracts, reviewers affirmed or revised their bids recorded in the conference management tool.

To quantitatively evaluate the performance of the SubSift tools, the bids made by reviewers were considered to be the ‘correct assignments’ against which SubSift’s automated assignments were compared. Disregarding the level of bid, a median of 88.2% of the papers recommended by SubSift were subsequently included in the reviewers’ own bids (precision). Furthermore, a median of 80.0% of the papers on which reviewers bid for were ones initially recommended to them by SubSift (recall).

These results suggest that the papers eventually bid on by reviewers were largely drawn from those that were assigned non-zero bids by SubSift. These results on real-world data in a practical setting are comparable with other published results using language models.

<sup>b</sup> ICML 2012 reviewing; <http://hunch.net/?p=2407>

score bias (do they tend to err on the accepting side or rather on the rejecting side?) and spread (do they tend to score more or less confidently?) we need a representative sample of papers with a reasonable distribution in quality. This is often problematic for single references as the number of papers  $m$  reviewed by a single reviewer is too small to be representative, and there can be considerable variation in the quality of papers among different batches that should not be attributed to reviewers. It is, however, possible to get more information about reviewer bias and confidence by leveraging the fact that papers are reviewed by several reviewers. For SIGKDD'09 we used a generative probabilistic model proposed by colleagues at Microsoft Research Cambridge with latent (unobserved) variables that can be inferred by message-passing techniques such as Expectation Propagation.<sup>35</sup> The latent variables include the true paper quality, the numerical score assigned by the reviewer, and the thresholds this particular reviewer uses to convert the numerical score to the observed recommendation on the seven-point scale. The calibration process is described in more detail in Flach et al.<sup>18</sup>

An interesting manifestation of reviewer variance came to light through an experiment with NIPS reviewing in 2014.<sup>27</sup> The PC chairs decided to have one-tenth (166) of the submitted papers reviewed twice, each by three reviewers and one area chair. It turned out the accept/reject recommendations of the two area chairs differed in about one quarter of the cases (43). Given an overall acceptance rate of 22.5%, roughly 38 of the 166 double-reviewed papers were accepted following the recommendation of one of the area chairs; about 22 of these would have been rejected if the recommendation of the other area chair had been followed instead (assuming the disagreements were uniformly distributed over the two possibilities), which suggests that more than half (57%) of the accepted papers would not have made it to the conference if reviewed a second time.

What can be concluded from what came to be known as the “NIPS experiment” beyond these basic numbers

is up for debate. It is worth pointing out that, while the peer review process eventually leads to a binary accept/reject decision, paper quality most certainly is not: while a certain fraction of papers clearly deserves to be accepted, and another fraction clearly deserves to be rejected, the remaining papers have pros and cons that can be weighed up in different ways. So if two reviewers assign different scores to papers this doesn't mean that one of them is wrong, but rather they picked up on different aspects of the paper in different ways.

We suggest a good way forward is to think of the reviewer's job as to “profile” the paper in terms of its strong and weak points, and separate the reviewing job proper from the eventual accept/reject decision. One could imagine a situation where a submitted paper could go to a number of venues (including the ‘null’ venue), and the reviewing task is to help decide which of these venues is the most appropriate one. This would turn the peer review process into a matching process, where publication venues have a distinct profile (whether it accepts theoretical or applied papers, whether it puts more value on novelty or on technical depth, among others) to be matched by the submission's profile as decided by the peer review process. Indeed, some conferences already have a separate journal track that implies some form of reviewing process to decide which venue is the most suitable one.<sup>c</sup>

### Assembling Peer Review Panels

The formation of a pool of reviewers, whether for conferences, journals, or funding competitions, is a non-trivial process that seeks to balance a range of objective and subjective factors. In practice, the actual process by which a program chair assembles a program committee varies from, at one extreme, inviting friends and co-authors plus their friends and co-authors, through to the other extreme of a formalized

election and representation mechanism. The current generation of CMSs do not offer computational support for the formation of a balanced program committee; they assume prior existence of the list of potential reviewers and instead concentrate on supporting the administrative workflow of issuing and accepting invitations.

**Expert finding.** This lack of tool support is surprising considering the body of relevant work in the long-established field of expert finding.<sup>2,11,15,34,47</sup> Over the years since the first Text Retrieval Conference (TREC) in 1992, the task of finding experts on a particular topic has featured regularly in this long-running conference series and is now an active subfield of the broader text information retrieval discipline. Expert finding has a degree of overlap with the fields of bibliometrics, the quantitative analysis of academic publications and other research-related literature,<sup>21,38</sup> and scientometrics, which extends the scope to include grants, patents, discoveries, data outputs and, in the U.K., more abstract concepts such as ‘impact.’<sup>15</sup> Expert finding tends to be more profile-based (for example, based on the text of documents) than link-based (for example, based on cross-references between documents) although content analysis is an active area of bibliometrics in particular and has been used in combination with citation properties to link research topics to specific authors.<sup>11</sup> Even though by comparison with bibliometrics, scientometrics encompasses additional measures, in practice the dominant approach in both domains is citation analysis of academic literature. Citation analysis measures the properties of networks of citation among publications and has much in common with hyperlink analysis on the Web, where these measures employ similar graph theoretic methods designed to model reputation, with notable examples including Hubs and Authorities, and PageRank. Citation graph analysis, using a particle-swarm algorithm, has been used to suggest potential reviewers for a paper on the premise that the subject of a paper is characterized by the authors it cites.<sup>39</sup>

Harvard's Profiles Research Network Software (RNS)<sup>d</sup> exploits both graph-based and text-based methods.


<sup>c</sup> For example, the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) has a journal track where accepted papers are presented at the conference but published either in the *Machine Learning* journal or in *Data Mining and Knowledge Discovery*.

<sup>d</sup> <http://profiles.catalyst.harvard.edu>


By mining high-quality bibliographic metadata from sources like PubMed, Profiles RNS infers implicit networks based on keywords, co-authors, department, location, and similar research. Researchers can also define their own explicit networks and curate their list of keywords and publications. Profiles RNS supports expert finding via a rich set of searching and browsing functions for traversing these networks. Profiles RNS is a noteworthy open source example of a growing body of research intelligence tools that compete to provide definitive databases of academics that, while varying in scope, scale and features, collectively constitute a valuable resource for a program chair seeking new reviewers. Well-known examples include free sites like academia.edu, Google Scholar, Mendeley, Microsoft Academic Search, ResearchGate, and numerous others that mine public data or solicit data directly from researchers themselves, as well as pay-to-use offerings like Elsevier's Reviewer Finder.

**Data issues.** There is a wealth of publicly available data about the expertise of researchers that could, in principle, be used to profile program committee members (without requiring them to choose keywords or upload papers) or to suggest a ranked list of candidate invitees for any given set of topics. Obvious data sources include academic home pages, online bibliographies, grant awards, job titles, research group membership, events attended as well as membership of professional bodies and other reviewer pools. Despite the availability of such data, there are a number of problems in using it for the purpose of finding an expert on a particular topic.

If the data is to be located and used automatically then it is necessary to identify the individual or individuals described by the data. Unfortunately a person's name is not guaranteed to be a unique identifier (UID): often not being globally unique in the first place, they can also be changed through title, choice, marriage, and so on. Matters are made worse because many academic reference styles use abbreviated forms of a name using initials. International variations in word ordering, character sets, and alternative spellings make name



**We suggest a good way is to think of a reviewer's job to "profile" the paper in terms of its strong and weak points, and separate the reviewing job proper from the eventual accept/reject decision.**



resolution even more challenging for a peer review tool. Indeed, the problem of author disambiguation is sufficiently challenging to have merited the investment of considerable research effort over the years, which has in turn led to practical tool development in areas with similar requirements to finding potential peer reviewers. For instance, Profiles RNS supports finding researchers with specific expertise and includes an Author Disambiguation Engine using factors such as name permutations, email address, institution affiliations, known co-authors, journal titles, subject areas, and keywords.

To address these problems in their own record systems, publishers and bibliographic databases like DBLP and Google Scholar have developed their own proprietary UID schemes for identifying contributors to published works. However, there is now considerable momentum behind the non-proprietary Open Researcher and Contributor ID (ORCID)<sup>e</sup> and publishers are increasingly mapping their own UIDs onto ORCID UIDs. A subtle problem remains for peer review tools when associating data, particularly academic publications, with an individual researcher because a great deal of academic work is attributed to multiple contributors. Hope for resolving individual contributions comes from a concerted effort to better document all outputs of research, including not only papers but also websites, datasets, and software, through richer metadata descriptions of Research Objects.<sup>10</sup>

**Balance and coverage.** Finding candidate reviewers is only part of a program chair's task in forming a committee—attention must also be paid to coverage and balance. It is important to ensure more popular areas get proportionately more coverage than less popular ones while also not excluding less well known but potentially important new areas. Thus, there is a subjective element to balance and coverage that is not entirely captured by the score matrix. Recent work seeks to address this for conferences by refining clusters, computed from a score matrix, using a form of crowdsourcing from the program committee and from the

<sup>e</sup> <http://orcid.org>

authors of accepted papers.<sup>1</sup> Another example of computational support for assembling a balanced set of reviewers comes not from conferences but from a U.S. funding agency, the National Science Foundation (NSF).

The NSF presides over a budget of over \$7.7 billion (FY 2016) and receives 40,000 proposals per year, with large competitions attracting 500–1,500 proposals; peer review is part of the NSF’s core business. Approximately a decade ago, the NSF developed Revaide, a data-mining tool to help them find proposal reviewers and to build panels with expertise appropriate to the subjects of received proposals.<sup>22</sup> In constructing profiles of potential reviewers the NSF decided against using bibliographic databases like Citeseer or Google Scholar, for the same reasons we discussed earlier. Instead they took a closed-world approach by restricting the set of potential reviewers to authors of past (single-author) proposals that had been judged ‘fundable’ by the review process. This ensured the availability of a UID for each author and reliable metadata, including the author’s name and institution, which facilitated conflict of interest detection. Reviewer profiles were constructed from the text of their past proposal documents (including references and résumés) as a vector of the top 20 terms with the highest tf-idf scores. Such documents were known to be all of similar length and style, which improved the relevance of the resultant tf-idf scores. The same is also true of the proposals to be reviewed and so profiles of the same type were constructed for these.

For a machine learning researcher, an obvious next step toward forming panels with appropriate coverage for the topics of the submissions would be to cluster the profiles of received proposals and use the resultant clusters as the basis for panels, for example, matching potential reviewers against a prototypical member of the cluster. Indeed, prior to Revaide the NSF had experimented with the use of automated clustering for panel formation but those attempts had proved unsuccessful for a number of reasons: the sizes of clusters tended to be uneven; clusters exhibited poor stability as new proposals arrived incrementally; there was a lack of alignment of pan-

els with the NSF organizational structure; and, similarly, no alignment with specific competition goals, such as increasing participation of underrepresented groups or creating results of interest to industry. So, eschewing clustering, Revaide instead supported the established manual process by annotating each proposal with its top 20 terms as a practical alternative to manually supplied keywords.

Other ideas for tool support in panel formation were considered. Inspired by conference peer review, NSF experimented with bidding but found that reviewers had strong preferences toward well-known researchers and this approach failed to ensure there were reviewers from all contributing disciplines of a multidisciplinary proposal—a particular concern for NSF. Again, manual processes won out. However, Revaide did find a valuable role for clustering techniques as a way of checking manual assignments of proposals to panels. To do this, Revaide calculated an “average” vector for each panel, by taking the central point of the vectors of its panel members, and then compared each proposal’s vector against every panel. If a proposal’s assigned panel is not its closest panel then the program director is warned. Using this method, Revaide proposed better assignments for 5% of all proposals. Using the same representation, Revaide was also used to classify orphaned proposals, suggesting a suitable panel. Although the classifier was only 80% accurate, which is clearly not good enough for a fully automated assignment, it played a valuable role within the NSF workflow: so, instead of each program director having to sift through, say, 1,000 orphaned proposals they received an initial assignment of, say, 100 of which they would need to reassign around 20 to other panels.

### Conclusion and Outlook

We have demonstrated that state-of-the-art tools from machine learning and artificial intelligence are making inroads to automate and improve parts of the peer review process. Allocating papers (or grant proposals) to reviewers is an area where much progress has been made. The combinatorial allocation problem can eas-

ily be solved once we have a score matrix assessing for each paper-reviewer pair how well they are matched.<sup>f</sup> We have described a range of techniques from information retrieval and machine learning that can produce such a score matrix. The notion of profiles (of reviewers as well as papers) is useful here as it turns a heterogeneous matching problem into a homogeneous one. Such profiles can be formulated against a fixed vocabulary (bag-of-words) or against a small set of topics. Although it is fashionable in machine learning to treat such topics as latent variables that can be learned from data, we have found stability issues with latent topic models (that is, adding a few documents to a collection can completely change the learned topics) and have started to experiment with handcrafted topics (for example, encyclopedia or Wikipedia entries) that extend keywords by allowing their own bag-of-words representations.

A perhaps less commonly studied area where nevertheless progress has been achieved concerns interpretation and calibration of the intermediate output of the peer reviewing process: the aspects of the reviews that feed into the decision making process. In their simplest form these are scores on an ordinal scale that are often simply averaged. However, averaging assessments from different assessors—which is common in other areas as well, for example, grading coursework—is fraught with difficulties as it makes the unrealistic assumption that each assessor scores on the same scale. It is possible to adjust for differences between individual reviewers, particularly when a reviewing history is available that spans multiple conferences. Such a global reviewing system that builds up persistent reviewer (and author) profiles is something that we support in principle, although many details need to be worked out before this is viable.

We also believe it would be beneficial if the role of individual reviewers shifted away from being an ersatz judge attempting to answer the ques-

<sup>f</sup> This holds for the simple version stated earlier, but further constraints might complicate the allocation problem.



tion “Would you accept this paper if it was entirely up to you?” toward a more constructive role of characterizing—and indeed, profiling—the paper under submission. Put differently, besides suggestions for improvement to the authors, the reviewers attempt to collect metadata about the paper that is used further down the pipeline to decide the most suitable publication venue. In principle, this would make it feasible to decouple the reviewing process from individual venues, something that would also enable better load balancing and scaling.<sup>46</sup> In such a system, authors and reviewers would be members of some central organization, which has the authority to assign papers to multiple publication venues—a futuristic scenario, perhaps, but it is worth thinking about the peculiar constraints that our current conference- and journal-driven system imposes, and which clearly leads to a sub-optimal situation in many respects.

The computational methods we described in this article have been used to support other academic processes outside of peer review, including a personalized conference planner app for delegates,<sup>5</sup> an organizational profiler<sup>36</sup> and a personalized course recommender for students based on their academic profile.<sup>41</sup> The accompanying table presented a few other possible future directions for computation support of academic peer review itself. We hope that they, along with this article, stimulate our readers to think about ways in which the academic peer review process—this strange dance in which we all participate in one way or another—can be future-proofed in a sustainable and scalable way. ■

g <http://www.subsift.com/ecmlpkdd2012/attendance/apps/>

## References

- André, P., Zhang, H., Kim, J., Chilton, L.B., Dow, S.P. and Miller, R.C. Community clustering: Leveraging an academic crowd to form coherent conference sessions. In *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing* (Palm Springs, CA, Nov. 7–9, 2013). B. Hartman and E. Horvitz, ed. AAAI, Palo Alto, CA.
- Balog, K., Azzopardi, L. and de Rijke, M. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th Annual International ACM Conference on Research and Development in Information Retrieval* (2006). ACM, New York, NY, 43–50.
- Benferhat, S. and Lang, J. Conference paper assignment. *International Journal of Intelligent Systems* 16, 10 (2001), 1183–1192.
- Blei, D.M., Ng, A.Y. and Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* (Mar. 2003), 993–1022.
- Bornmann, L., Bowman, B., Bauer, J., Marx, W., Schier, H. and Palzenberger, M. Standards for using bibliometrics in the evaluation of research institutes. *Next Generation Metrics*, 2013.
- Boxwala, A.A., Dierks, M., Keenan, M., Jackson, S., Hanscom, R., Bates, D.W. and Sato, L. Review paper: Organization and representation of patient safety data: Current status and issues around generalizability and scalability. *J. American Medical Informatics Association* 11, 6 (2004), 468–478.
- Brixy, J., Johnson, T. and Zhang, J. Evaluating a medical error taxonomy. In *Proceedings of the American Medical Informatics Association Symposium*, 2002.
- Charlin, L. and Zemel, R. The Toronto paper matching system: An automated paper-reviewer assignment system. In *Proceedings of ICM Workshop on Peer Reviewing and Publishing Models*, 2013.
- Charlin, L., Zemel, R. and Boutillier, C. A framework for optimizing paper matching. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence* (Corvallis, OR, 2011). AUAI Press, 86–95.
- De Roure, D. Towards computational research objects. In *Proceedings of the 1st International Workshop on Digital Preservation of Research Methods and Artefacts* (2013). ACM, New York, NY, 16–19.
- Deng, H., King, I. and Lyu, M.R. Formal models for expert finding on DBLP bibliography data. In *Proceedings of the 8th IEEE International Conference on Data Mining (2008)*. IEEE Computer Society, Washington, D.C., 163–172.
- Devedzić, V. Understanding ontological engineering. *Commun. ACM* 45, 4 (Apr. 2002), 136–144.
- Di Mauro, N., Basile, T. and Ferilli, S. Grape: An expert review assignment component for scientific conference management systems. *Innovations in Applied Artificial Intelligence. LNCS 3533* (2005). M. Ali and F. Esposito, eds. Springer, Berlin Heidelberg, 789–798.
- Dumais, S.T. and Nielsen, J. Automating the assignment of submitted manuscripts to reviewers. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (1992). ACM, New York, NY, 233–244.
- Fang, H. and Zhai, C. Probabilistic models for expert finding. In *Proceedings of the 29th European Conference on IR Research* (2007). Springer-Verlag, Berlin, Heidelberg, 418–430.
- Ferilli, S., Di Mauro, N., Basile, T., Esposito, F. and Biba, M. Automatic topics identification for reviewer assignment. *Advances in Applied Artificial Intelligence. LNCS 4031* (2006). M. Ali and R. Dapogny, eds. Springer, Berlin Heidelberg, 721–730.
- Flach, P. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge University Press, 2012.
- Flach, P.A., Spiegler, S., Golénia, B., Price, S., Herbrich, J.G.R., Graepel, T. and Zaki, M.J. Novel tools to streamline the conference review process: Experiences from SIGKDD'09. *SIGKDD Explorations* 11, 2 (Dec. 2009), 63–67.
- Garg, N., Kavitha, T., Kumar, A., Mehlhorn, K., and Mestre, J. Assigning papers to referees. *Algorithmica* 58, 1 (Sept. 2010), 119–136.
- Goldsmith, J. and Sloan, R.H. The AI conference paper assignment problem. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence* (2007).
- Harnad, S. Open access scientometrics and the U.K. research assessment exercise. *Scientometrics* 79, 1 (Apr. 2009), 147–156.
- Hettich, S. and Pazzani, M.J. Mining for proposal reviewers: Lessons learned at the National Science Foundation. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2006). ACM, New York, NY, 862–871.
- Jennings, C. Quality and value: The true purpose of peer review. *Nature*, 2006.
- Karimzadehgan, M. and Zhai, C. Integer linear programming for constrained multi-aspect committee review assignment. *Inf. Process. Manage.* 48, 4 (July 2012), 725–740.
- Karimzadehgan, M., Zhai, C. and Belford, G. Multi-aspect expertise matching for review assignment. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (2008). ACM, New York, NY 1113–1122.
- Kou, N.M., U, L.H., Mamoulis, N. and Gong, Z. Weighted coverage based reviewer assignment. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM, New York, NY, 2031–2046.
- Langford, J. and Guzdial, M. The arbitrariness of reviews, and advice for school administrators. *Commun. ACM* 58, 4 (Apr. 2015), 12–13.
- Lawrence, P.A. The politics of publication. *Nature* 422 (Mar. 2003), 259–261.
- Ley, M. The DBLP computer science bibliography: Evolution, research issues, perspectives. In *Proceedings of the 9th International Symposium on String Processing and Information Retrieval* (London, U.K., 2002). Springer-Verlag, 1–10.
- Liu, X., Suel, T. and Memon, N. A robust model for paper reviewer assignment. In *Proceedings of the 8th ACM Conference on Recommender Systems* (2014). ACM, New York, NY, 25–32.
- Long, C., Wong, R.C., Peng, Y. and Ye, L. On good and fair paper-reviewer assignment. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining* (Dallas, TX, Dec. 7–10, 2013), 1145–1150.
- Mehlhorn, K., Vardi, M.Y. and Herbrich, M. Publication culture in computing research (Dagstuhl Perspectives Workshop 12452). *Dagstuhl Reports* 2, 11 (2013).
- Meyer, B., Choppy, C., Staunstrup, J. and van Leeuwen, J. Viewpoint: Research evaluation for computer science. *Commun. ACM* 52, 4 (Apr. 2009), 31–34.
- Mimno, D. and McCallum, A. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 2007, 500–509.
- Minka, T. Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*. J.S. Breese and D. Koller, Eds. Morgan Kaufmann, 2001, 362–369.
- Price, S. and Flach, P.A. Mining and mapping the research landscape. In *Proceedings of the Digital Research Conference*. University of Oxford, Sept. 2013.
- Price, S., Flach, P.A., Spiegler, S., Bailey, C. and Rogers, N. SubSift Web services and workflows for profiling and comparing scientists and their published works. *Future Generation Comp. Syst.* 29, 2 (2013), 569–581.
- Pritchard, A. et al. Statistical bibliography or bibliometrics. *J. Documentation* 25, 4 (1969), 348–349.
- Rodriguez, M.A. and Bollen, J. An algorithm to determine peer-reviewers. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. ACM, New York, NY, 319–328.
- Sidiropoulos, N.D. and Tsakonas, E. Signal processing and optimization tools for conference review and session assignment. *IEEE Signal Process. Mag.* 32, 3 (2015), 141–155.
- Surpatean, A., Smirnov, E.N. and Manie, N. Master orientation tool. *ECAI 242, Frontiers in Artificial Intelligence and Applications*. L. De Raedt, C. Bessière, D. Dubois, P. Doherty, P. Frasconi, F. Heintz, and P.J.F. Lucas, Eds. IOS Press, 2012, 995–996.
- Tang, W., Tang, J., Lei, T., Tan, C., Gao, B. and Li, T. On optimization of expertise matching with various constraints. *Neurocomputing* 76, 1 (Jan. 2012), 71–83.
- Tang, W., Tang, J. and Tan, C. Expertise matching via constraint-based optimization. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (Vol. 1). IEEE Computer Society, Washington, DC, 2010, 34–41.
- Taylor, C.J. On the optimal assignment of conference papers to reviewers. Technical Report MS-CIS-08-30, Computer and Information Science Department, University of Pennsylvania, 2008.
- Terry, D. Publish now, judge later. *Commun. ACM* 57, 1 (Jan. 2014), 44–46.
- Vardi, M.Y. Scalable conferences. *Commun. ACM* 57, 1 (Jan. 2014), 5.
- Yimam-Seid, D. and Kobsa, A. Expert finding systems for organizations: Problem and domain analysis and the DEMOIR approach. *J. Organizational Computing and Electronic Commerce* 13 (2003).

**Simon Price** (simon.price@bristol.ac.uk) is a Visiting Fellow in the Department of Computer Science at the University of Bristol, U.K., and a data scientist at Cag Gemini.

**Peter A. Flach** (peter.flach@bristol.ac.uk) is a professor of artificial intelligence in the Department of Computer Science at the University of Bristol, U.K. and editor-in-chief of the *Machine Learning* journal.

Copyright held by owners/authors.



Watch the authors discuss their work in this exclusive *Communications* video. <http://cacm.acm.org/videos/computational-support-for-academic-peer-review>



# Attention: Undergraduate *and* Graduate Computing Students

There's an **ACM Student Research Competition** at a SIG Conference of interest to you!

It's hard to put the **ACM Student Research Competition (SRC)** experience into words, but we'll try...



"The ACM SRC has been an extraordinary opportunity to expose my research and get feedback from an audience of experts in my field. What makes SRC a unique and integral experience is that along the phases of the competition, not only the research work itself is evaluated but also the soft skills of the participants."

**Miguel Angel Aguilar**  
RWTH Aachen University | PACT 2015



"The ACM SRC is an excellent platform to present your ongoing research and get feedback from experienced people in the community. It helps you improve your research and presentation skills, and lets you gauge the interest of the community in your ongoing research work."

**Swarnendu Biswas**  
Ohio State University | PLDI 2015

"The SPLASH 2015 SRC presented an opportunity to experience work in software engineering and to practice presentation and discussion skills. Conversations with fellow students and researchers about opportunities in the field will guide my future for years to come."

**Andrew Kofink**  
North Carolina State University | SPLASH 2015



"Participating in the SRC was an amazing opportunity. It was my first time attending any conference, and it really showed me how to pitch my research project, and interact with other researchers. I will carry this experience with me in my future academic and professional endeavors."

**Michele Hu**  
Cornell Tech | ASSETS 2015



"SRC provided me a unique opportunity of showcasing early stage research work to the pioneers of the community. The constructive feedback and excitement received during SRC later led to a broader and stronger piece of future work. Since it's organized in a competitive manner and judged by experts, it helps one in both the technical as well as communication fronts."

**Puneet Jain**  
Duke University | MobiCom 2015



"The ACM SRC was a fantastic opportunity to get my research in front of expert researchers who would not otherwise have seen it. I received invaluable feedback on both my research direction and my presentation and communication skills that will serve me well going forward. Later in my career, I hope I have the opportunity to participate in future SRCs as a judge and mentor to pay it forward to the generation after mine."

**Christopher Theisen**  
North Carolina State University | FSE 2015

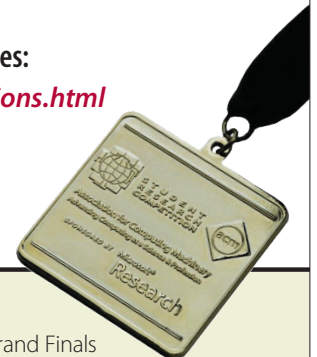
"ACM SRC gives students an incentive to think about questions beyond the daily focus: Why am I doing this research? How does it impact society? Why should other researchers care? Which future directions to pursue? Eventually, each researcher will need to answer these questions in their career, and it is helpful to face them early in the SRC."

**Ivan Ruchkin**  
Carnegie Mellon University | MODELS 2015



Check the SRC Submission Dates:

<http://src.acm.org/submissions.html>



- ◆ Participants receive: \$500 (USD) travel expenses
- ◆ All Winners receive a handsome medal and monetary award. First place winners advance to the SRC Grand Finals
- ◆ Grand Finals Winners receive a handsome certificate and monetary award at the ACM Awards Banquet

**Questions?** Contact Nanette Hernandez, ACM's SRC Coordinator: [hernandez@hq.acm.org](mailto:hernandez@hq.acm.org)

# research highlights

---

P. 82

**Technical  
Perspective**  
**The Power of Wi-Fi  
to Deliver Power**

By Srinivasan Keshav

P. 83

## Powering the Next Billion Devices with Wi-Fi

By Vamsi Talla, Bryce Kellogg, Benjamin Ransford,  
Saman Naderiparizi, Joshua R. Smith, and Shyamnath Gollakota

---

P. 92

**Technical  
Perspective**  
**Data Distribution  
for Fast Joins**

By Leonid Libkin

P. 93

## Reasoning on Data Partitioning for Single-Round Multi-Join Evaluation in Massively Parallel Systems

By Tom J. Ameloot, Gaetano Geck, Bas Ketsman,  
Frank Neven, and Thomas Schwentick

# Technical Perspective

## The Power of Wi-Fi to Deliver Power

By Srinivasan Keshav

IN THE LATE 1990s, the rapidly dropping costs of microprocessors, of wireless network interfaces, and of sensors led some researchers to propose a powerful vision: a vision that tiny, wirelessly connected, computerized sensors could be scattered about like grains of smart dust and that these ‘motes’ would self-organize into a network that would allow us to weave intelligence into the physical world.<sup>1,2</sup> This would allow us to intelligently control a diverse array of physical systems, making them more efficient, less power hungry, and more responsive to human needs. For example, we could reduce the costs of heating and lighting a building, providing these services only to occupied rooms; could measure every tremor in an earthquake zone, predicting large quakes; or could let computers sense blood sugar levels and control insulin pumps, making life more pleasant for diabetics. Thousands of researchers were inspired by this vision to work on many aspects of these ‘wireless sensor networks,’ making this a rich field of scientific inquiry.

However, one aspect of wireless sensor networks has detracted us from realizing this powerful vision. This is the need to provide power to sensor motes. It has turned out that powering a sensor using batteries makes them large, expensive, and unwieldy. Instead of scattering them to the winds, they need to be very carefully sited, so that batteries can be replaced from time to time, and so that energy would not be wasted on expensive wireless packet transmissions. This has greatly reduced the scope of wireless sensor networks, making them more of a niche technology than one would otherwise expect.

Despite this setback, a small group of researchers have held true to the original vision. Their line of attack has been to use *energy harvesting*, that is, gathering energy from the environ-


ment. Approaches include using tiny photovoltaic panels to harvest light, piezoelectric crystals to harvest vibrations, and antennas to harvest microscopic amounts of energy from radio and TV signals. However, these approaches have had limited success because sensors that rely on energy harvesting alone cannot be guaranteed to receive energy when they need it: they may be in the dark, in vibration-free environments, or in remote areas with a quiescent electromagnetic spectrum.

In the following paper by Talla et al., the authors turn the problem on its head. Instead of focusing on energy harvesting, they focus on wireless energy transfer. In their approach, a sensor mote harvests energy wirelessly transmitted by a dedicated power supplier. By itself, this is not particularly novel, in that this has been used by Radio Frequency Identification (RFID) systems for many years. What is clever about the paper is the authors use ubiquitous Wi-Fi devices both to supply and to harvest radio frequency energy. More specifically, they modify standard Wi-Fi chipsets to transmit special power packets that can be used to deliver power to a mote. Moreover, to prevent the energy harvester on a mote (a capacitor coupled to an an-

tenna) from losing energy due to self-discharge, they send power packets on multiple Wi-Fi channels. They also use a cleverly chosen power packet transmission schedule to ensure that power packets minimally affect data transmission to other nodes.

The net result is their system, whose transmitter uses a stock Wi-Fi chipset, and whose receiver uses custom hardware, can wireless power sensors, such as a camera and a temperature probe. They can also wireless trickle charge a standard battery over the air.

While this is a big step toward the original vision of wireless sensor networks, unfortunately, the ultimate vision is still not within grasp. Wi-Fi antennas are several centimeters long, which makes the sensors not quite dust-like. The range over which power can be transmitted using this approach is also fairly small, less than a few meters. Moreover, sensors powered in this way can only operate at tens of Hz, at best. No doubt, these limitations will be overcome in years to come.

Although this paper does not deliver on the original vision of wireless sensor networks, it is nevertheless well worth reading, if only as an exercise in lateral thinking. It challenges our customary view of Wi-Fi as a data transmission technology, and shows that Wi-Fi can be used to deliver power as well, a rather surprising observation. Maybe it will stimulate you too to use standard technologies in unconventional ways? 

**What is clever about the following paper is the authors use ubiquitous Wi-Fi devices both to supply and harvest radio frequency energy.**

### References

1. Estrin, D. et al. Instrumenting the world with wireless sensor networks. In *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4.
2. Kahn, J.M., Katz, R.H. and Pister, K.S.J. Next century challenges: Mobile networking for ‘Smart Dust.’ In *Proceedings of the 5th annual ACM/IEEE International Conference on Mobile Computing and Networking*. ACM, 1999.

Srinivasan Keshav is a professor and Cisco Chair at the David R. Cheriton School of Computer Science at the University of Waterloo, Canada.

Copyright held by author.

# Powering the Next Billion Devices with Wi-Fi

By Vamsi Talla, Bryce Kellogg, Benjamin Ransford, Saman Naderiparizi, Joshua R. Smith, and Shyamnath Gollakota

## Abstract

We present the first *power over Wi-Fi* system that delivers power to low-power sensors and devices and works with existing Wi-Fi chipsets. We show that a ubiquitous part of wireless communication infrastructure, the Wi-Fi router, can provide far field wireless power without significantly compromising the network's communication performance. Building on our design, we prototype battery-free temperature and camera sensors that we power with Wi-Fi at ranges of 20 and 17 ft, respectively. We also demonstrate the ability to wirelessly trickle-charge nickel-metal hydride and lithium-ion coin-cell batteries at distances of up to 28 ft. We deploy our system in six homes in a metropolitan area and show that it can successfully deliver power via Wi-Fi under real-world network conditions without significantly degrading network performance.

## 1. INTRODUCTION

In the late 19th century, Nikola Tesla dreamed of eliminating wires for both power and communication.<sup>16</sup> As of the early 21st century, wireless communication is extremely well established—billions of people rely on it every day. Wireless power, however, has not been as successful. In recent years, near-field, short range schemes have gained traction for certain range-limited applications, like powering implanted medical devices<sup>20</sup> and recharging cars<sup>3</sup> and phones from power delivery mats.<sup>8</sup> More recently, researchers have demonstrated the feasibility of powering sensors and devices in the far field using RF signals from TV<sup>7</sup> and cellular<sup>19</sup> base stations. This is exciting, because in addition to enabling power delivery at farther distances, RF signals can simultaneously charge multiple sensors and devices because of their broadcast nature.

In this work, we show that a ubiquitous part of wireless infrastructure, the Wi-Fi router, can be used to provide far-field wireless power without significantly compromising network performance. This is attractive for three key reasons:

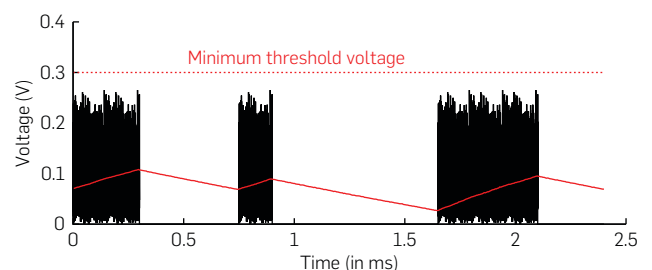
- Unlike TV and cellular transmissions, Wi-Fi is ubiquitous in indoor environments and operates in unlicensed spectrum (the “ISM” band) where transmissions can legally be optimized for power delivery. Repurposing Wi-Fi networks for power delivery can ease the deployment of RF-powered devices without additional power infrastructure.
- Wi-Fi uses OFDM, an efficient waveform for power delivery because of its high peak-to-average power ratio.<sup>17</sup> Given Wi-Fi's economies of scale, Wi-Fi chipsets provide a cheap platform for sending these power-optimized waveforms, enabling efficient power delivery.

- Sensors and mobile devices are increasingly equipped with 2.4 GHz antennas for communication via Wi-Fi, Bluetooth, or ZigBee. We can, in principle, use the same antenna for both communication and Wi-Fi power harvesting with a negligible effect on device size.

The key challenge for power delivery over Wi-Fi is the fundamental mismatch between the requirements for power delivery and the Wi-Fi protocol. To illustrate this, Figure 1 plots the voltage at a tuned harvester in the presence of Wi-Fi transmissions. While the harvester can gather energy during Wi-Fi transmissions, the energy leaks during silent periods. In this case, the Wi-Fi transmissions cannot meet the platform's minimum voltage requirement. Unfortunately for power delivery, silent periods are inherent to a distributed medium access protocol such as Wi-Fi, in which multiple devices share the same wireless medium. Continuous transmission from the router would be optimal for power delivery but would significantly degrade the performance of Wi-Fi clients and other nearby Wi-Fi networks.

This paper introduces *PoWiFi*, the first power over Wi-Fi system that delivers power to energy-harvesting sensors and devices while preserving network performance. We achieve this by codesigning harvesting hardware circuits and Wi-Fi router transmissions. At a high level, a router running PoWiFi imitates a continuous transmission from a harvester's perspective while minimizing the impact on Wi-Fi clients and neighboring Wi-Fi networks. The key intuition is that it is unlikely that all the Wi-Fi channels are simultaneously occupied at the same instant. PoWiFi opportunistically injects superfluous broadcast traffic (which we call *power packets*)

**Figure 1. Key challenge with Wi-Fi power delivery. While the harvester can gather power during Wi-Fi transmissions, the power leaks during silent periods, limiting Wi-Fi's ability to meet the minimum voltage requirements of the hardware.**



The original version of this paper was published in *ACM CoNext 2015*.

on nonoverlapping Wi-Fi channels to maximize the *cumulative* occupancy across the channels. To harvest this energy, we introduce the first multichannel harvester that efficiently harvests power across multiple Wi-Fi channels and generates the 1.8–2.4V necessary to run microcontrollers and sensor systems.

To be practical, PoWiFi must not significantly degrade network performance. So our second component is a transmission mechanism that minimizes the impact on Wi-Fi performance while effectively providing continuous power to harvesters. Specifically, to minimize the impact on associated Wi-Fi clients, PoWiFi injects power packets on a channel only when the number of data packets queued at the Wi-Fi interface is below a threshold. Further, the router transmits power packets at the highest Wi-Fi bit rates to minimize their duration, maximizing fairness to other Wi-Fi transmitters.

To further minimize its impact on neighboring Wi-Fi networks, PoWiFi uses two key techniques.

- *Rectifier-aware transmissions.* The key intuition is that when there are packets on the air, a harvester's temporary energy supply charges exponentially, but it also discharges exponentially during silent periods. To balance power delivery and channel occupancy, PoWiFi must minimize energy loss due to leakage. We achieve this by designing an occupancy modulation scheme that jointly optimizes the rectifier's voltage behavior and the Wi-Fi network's throughput to ensure that harvesting sensors can meet their duty-cycling requirements (see Rectifier-aware PoWiFi transmissions section).
- *Scalable concurrent transmissions.* A key goal is to maintain good network performance when there are multiple PoWiFi routers in an area. Our insight is that PoWiFi's power packets do not contain useful data, and so the transmissions from multiple PoWiFi routers can safely collide. Further, by making each PoWiFi router transmit random power packets, we ensure that concurrent packet transmissions do not destructively interfere to reduce available power at sensors.

We build prototype PoWiFi routers using Atheros chipsets and harvesters using off-the-shelf components. Our experiments demonstrate the following:

- The power packets at the PoWiFi router do not noticeably affect TCP or UDP throughput or webpage load times<sup>1</sup> at

an associated client. Meanwhile, PoWiFi achieves an average cumulative occupancy of 95.4% across the three nonoverlapping 2.4 GHz Wi-Fi channels.

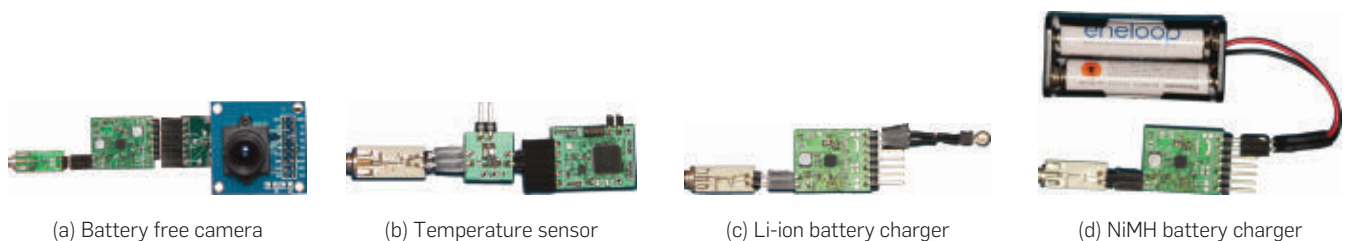
- PoWiFi's unintrusive transmission strategy allows neighboring Wi-Fi networks to achieve better-than-equal-share fairness, because a PoWiFi router transmits power packets at the highest bit rate to minimize its channel occupancy.
- Using a rectifier-aware transmission scheme that can adapt to a harvester's energy needs, PoWiFi's per-channel occupancy is as low as 4.4% while delivering power to a sensor 16 ft away that reads temperature values once every minute.
- We perform a proof-of-concept evaluation of our concurrent transmission mechanism with one, three, and six PoWiFi routers. While the variance of neighboring Wi-Fi networks' throughput slightly increases, their mean throughput does not statistically differ. This shows the feasibility of scaling our design with multiple PoWiFi routers.

To demonstrate the potential of our design, we build two battery-free, Wi-Fi powered sensing systems shown in Figure 2: a temperature sensor and a camera. The devices use Wi-Fi power to run their sensors and a programmable microcontroller that collects the data and sends it over a UART interface. The camera and temperature-sensor prototypes can operate battery-free at distances of up to 17 and 20 ft, respectively, from a PoWiFi router. As expected, the duty cycle at which these sensors can operate decreases with distance. Further, the sensors can operate in through-the-wall scenarios when separated from the router by various wall materials.

We also integrate our harvester with 2.4V nickel-metal hydride (NiMH) and 3.0V lithium-ion (Li-Ion) coin-cell batteries. We build battery-recharging versions of the above sensors wherein PoWiFi trickle charges the batteries. The battery-recharging sensors can run energy-neutral operations at distances of up to 28 ft.

Finally, we deploy PoWiFi routers in six homes in a metropolitan area. Each home's occupants used the PoWiFi router for their Internet access for 24 h. Even under real-world network conditions, PoWiFi efficiently delivers power while having a minimal impact on user experience.

**Figure 2. Prototype hardware demonstrating PoWiFi's potential. The prototypes harvest energy from Wi-Fi signals through a standard 2 dBi Wi-Fi antenna (not shown). The low gain antenna ensures that the device is agnostic to the antenna orientation and placement. We developed (a) a battery free camera to capture images, (b) a temperature sensor to measure temperature, (c) a Li-ion battery charger, and (d) a NiMH battery charger.**



## 1.1. Limitations

Given today's FCC 1 W limit on transmitters in the ISM band that Wi-Fi uses, power over Wi-Fi is limited to low-power sensors and devices and cannot, for example, recharge smartphones that require 5 W. Further, the range of our system is determined by the sensitivity of our harvester hardware, which is built with off-the-shelf components. We believe that an ASIC design would be able to improve the harvester's sensitivity and double PoWiFi's power-delivery range. Finally, while our current design uses a single antenna, in principle we can use multiple antennas to focus more power toward a sensor and increase the range, but such optimizations are beyond the scope of this paper.

## 2. UNDERSTANDING WI-FI POWER

To understand the ability of a Wi-Fi router to deliver power, we run experiments with our organization's Asus RT-AC68U router and a temperature sensor. The router operates on Channel 5 and is set to transmit 23 dBm power on each of its three 4.04 dBi antennas. The temperature sensor is battery free and uses our RF harvester to draw power from Wi-Fi signals. An RF harvester is a device that converts incoming alternating current (AC) radio signals into direct current (DC). A typical RF harvester consists of two stages: a rectifier that converts the incoming radio signal oscillating at 2.4 GHz into DC voltage, and a DC-DC converter that boosts this voltage to a higher value. Every sensor or microcontroller requires a minimum voltage to run meaningful operations and the DC-DC converter ensures that these requirements are met. The key limitation in harvesting power is that every DC-DC converter has a minimum input voltage threshold below which it cannot operate. We use the DC-DC converter with the lowest threshold of 300 mV.<sup>12</sup>

We place the sensor 10 ft from the router for 24 h and measure the voltage at the rectifier's output throughout our experiments. We also capture the packet transmissions from the router using a high-frequency oscilloscope connected through a splitter. Over the tested period, the sensor did not reach the 300 mV threshold. Figure 1 plots both the packet transmissions and the rectifier voltage during a period of peak network utilization. It shows that while the sensor can harvest energy during the Wi-Fi packet transmission, there is no input power during the silent slots. The energy leakages during these periods ensure that the voltage does not cross the 300 mV threshold.

## 3. PoWiFi

PoWiFi combines two elements: (1) a Wi-Fi transmission strategy that delivers power on multiple Wi-Fi channels and (2) energy-harvesting hardware that can efficiently harvest from multiple Wi-Fi channels simultaneously. See the companion technical report<sup>14</sup> for details on the design of the energy-harvesting hardware.

### 3.1. PoWiFi router design

Our key insight is that, at any moment, it is unlikely that all Wi-Fi channels will be occupied. Thus, PoWiFi opportunistically injects power packets across multiple Wi-Fi channels with a goal of maximizing *cumulative* occupancy.

Specifically, it injects 1500-byte UDP broadcast datagrams with a 100 us inter-packet delay at the highest 802.11g bit rate of 54 Mbps on the three nonoverlapping 2.4 GHz Wi-Fi channels (1, 6, and 11). A PoWiFi router enqueues these broadcast packets only when the number of frames in the wireless interface's transmit queue is below a threshold (five frames). If the queue's depth is at or above this threshold, then there are already enough power and Wi-Fi client packets in the queue to maximize channel occupancy.

PoWiFi must also provide fairness to traffic from nearby networks. Since the PoWiFi router performs carrier sensing and transmits broadcast packets at the highest 802.11g bit rate, its individual frames are as short and unintrusive as possible. PoWiFi thereby provides better-than-equal-share fairness for transmissions from other networks. The rest of this section describes two techniques that further reduce PoWiFi's effect on neighboring networks.

**Rectifier-aware PoWiFi transmissions.** When a PoWiFi transmitter knows a harvester's electrical characteristics, it can tune its transmission strategy to precisely fit the device's power requirements. For example, suppose we need to read a temperature sensor once per minute. PoWiFi can modulate its occupancy to deliver energy to the harvester so that the sensor reaches its required voltage of 2.4 V just in time, minimizing the total channel occupancy subject to this goal and thereby minimizing its effect on other networks.

*Empirically modeling rectifier voltage.* A rectifier converts incoming Wi-Fi transmissions into DC voltage to charge a storage capacitor. Once the voltage on the capacitor reaches the required threshold ( $V_{th} = 2.4$  V for the temperature sensor), a reading occurs. Suppose the average power at the harvester after multipath reflections and attenuation is  $P_{in}$  and the channel occupancy of the PoWiFi router packets is  $C$ . To a first approximation, the harvester's behavior can be modeled as a DC voltage source charging a capacitor through a resistor. The difference, however, is that the approximated resistance value depends on the impedance of the harvester's diodes, which is a function of  $P_{in}$  and  $C$ . We can write the voltage as a function of time as

$$V(t) = V_0 e^{-t/\tau(P_{in}, C)} + V_{max}(P_{in}, C) \left(1 - e^{-t/\tau(P_{in}, C)}\right),$$

where  $V_0$  is the initial voltage,  $\tau$  is the time constant, and  $V_{max}$  is the maximum achievable voltage. Note that both  $\tau$  and  $V_{max}$  are functions of  $P_{in}$  and the channel occupancy.

Given the nonlinearities of diodes, it is difficult to obtain closed-form solutions for  $\tau(P_{in}, C)$  and  $V_{max}(P_{in}, C)$ . We instead connected the harvester through a cabled setup to a Wi-Fi source with variable input power and channel occupancy and measured the output voltage. We fitted the resulting data with the proposed exponential model to estimate how  $\tau$  and  $V_{max}$  vary with input power and channel occupancy. The key properties of our model fitting are: (1)  $V_{max}$  is inverse-linearly proportional to the input power and channel occupancy; (2) the time constant  $\tau$  is exponentially proportional to the input power and/or the channel occupancy; and (3) it takes exponentially more time for the same increment in the voltage at a higher voltage value than at a lower one.

We next describe how PoWiFi can modulate its channel occupancy using this empirical model, while minimizing its effect on neighboring Wi-Fi networks.

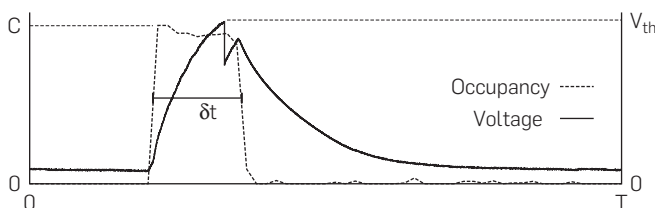
*Joint optimization for efficient power delivery.* To reduce the impact of power packets on neighboring Wi-Fi networks, PoWiFi must minimize the total number of power packets required to collect a sensor reading. Our key intuition is that when there are packets on the air, the capacitor charges exponentially. However, when there are no packets, the voltage on the capacitor discharges exponentially. To maximize the effectiveness of power delivery, PoWiFi must minimize capacitor leakage. We achieve this by using the channel-occupancy modulation scheme described above and shown in Figure 3. In every sensor update time window ( $T$ ), the router transmits no power packets for a period ( $T - \delta t$ ), then transmits power packets for a period of  $\delta t$ , targeting a channel occupancy of  $0 < C \leq 1$ . When the channel occupancy is zero, the voltage on the capacitor is very low and there is no leakage. However, when a sensor update is required, a high channel occupancy continuously charges the capacitor (minimizing leakage) to maximize the effectiveness of power delivery. Our goal is to find  $\delta t$  and  $C$  to minimize the mean of the power packet occupancy given by  $C * \frac{\delta t}{T}$ .

We find these values by substituting different  $C$  and  $\delta t$  in our empirical model and computing the minimum value. We reduce the search space by noting that for a given  $P_{in}$ , there is a minimum value of  $C$  below which the threshold voltage is not achievable. Further, given a channel occupancy, we know the time constant that limits  $\delta t$  to a maximum value of  $\tau(P_{in}, C)$ . Finally, we limit the granularity by which channel occupancy can be modulated to 10%. Using these values we reduce the search space to 75 points.

We note two main points. First, the above description assumes that the router can estimate the available power,  $P_{in}$ , at the sensor. To bootstrap this value, PoWiFi initially transmits power packets at a high occupancy of around 90% and notes the times when the sensor outputs a reading. PoWiFi uses our empirical model to estimate  $P_{in}$  for the next cycle. At the end of every cycle it re-estimates  $P_{in}$  to account for wireless channel changes. Second, in the presence of multiple sensors, we can optimize the parameters to satisfy the minimum duty-cycle requirement across all the sensors, but we omit this simple extension for brevity.

**Scaling with concurrent PoWiFi transmissions.** A practical

**Figure 3. Rectifier-aware power Wi-Fi transmissions and corresponding rectifier voltages. The plot shows the optimized rectifier-aware PoWiFi transmission and the corresponding voltage at the storage capacitor.  $V_{in} = 2.4V$  and  $\delta t = 10s$  for a temperature sensor reading every minute at the maximum operating distance.**



issue with each PoWiFi router independently introducing power packets is that such a system would not preserve network performance in the presence of many PoWiFi routers. Useful Wi-Fi capacity would degrade at least linearly with the number of PoWiFi routers.

To address this scaling problem, we enable concurrent transmissions from PoWiFi routers that are in decoding range of one another. Our key insight is that since power packets do not contain useful data, transmissions from multiple PoWiFi routers can safely collide. Further, if each PoWiFi router transmits a random power packet, we can ensure that concurrent packet transmissions do not destructively interfere to reduce the power available to harvesters.

Specifically, in our system, we have a leader PoWiFi router that transmits the energy pattern as shown in Figure 4. The pattern consists of a short packet with a 1-byte payload transmitted at 54Mbps, followed by a Distributed Interframe Space (DIFS) period and then a power packet. Other PoWiFi routers decode this short packet and join the packet transmission of the leader router within the DIFS period. This strategy ensures that all nearby PoWiFi routers transmit power packets concurrently and hence do not reduce the Wi-Fi network's capacity.

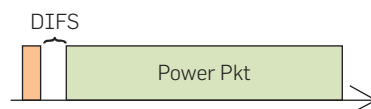
As in previous work that used concurrent transmissions,<sup>6</sup> we enable follower routers to transmit simultaneously in software by adjusting contention-window and noise-floor parameters to prevent carrier-sense backoff, and by placing power packets in the high-priority queue. However, PoWiFi could not turnaround and begin transmission within from the software layer within a DIFS duration; we believe that with better access to the router's hardware queues, PoWiFi could turnaround within a DIFS period. Further, one can design distributed algorithms to find the leader router whose transmissions can be decoded by all other PoWiFi routers, but we consider this to be outside the scope of this paper.

#### 4. EVALUATION

We build our harvester prototypes using commercial off-the-shelf components on printed circuit boards. We implement PoWiFi routers using three Atheros AR9580 chipsets that independently run the algorithm in Section 3.1 on channels 1, 6, and 11. The chipsets are connected via amplifiers to 6 dBi Wi-Fi antennas separated by 6.5 cm. Our prototype router provides Internet access to its associated clients on channel 1 via NAT and transmits at 30 dBm, the FCC limit for power in the ISM band. All our sensor and harvester benchmark evaluations were performed in a busy office network where the average cumulative occupancy across the three channels was about 90%.

Both power and data packets contribute to our router's

**Figure 4. Energy pattern for concurrent power packet transmissions. It consists of a short packet with a 1-byte payload transmitted at 54Mbps, followed by a DIFS period and a power packet transmission.**





channel occupancy. To measure occupancy, we use `aircrack-ng`'s `airmon-ng` tool to add a monitor interface to each of the router's active wireless interfaces. Then, on each monitor interface, we start `tcpdump` to record the radiotap headers for all frames and their retransmissions. We use `tshark` to extract frames sent by the router, recording the corresponding bitrate and frame size (in bytes). We then compute the average channel occupancy as  $\sum_{i \in \text{frames}} \frac{\text{size}_i}{\text{rate}_i \times \text{total\_duration}}$ .

#### 4.1. Effect on Wi-Fi clients

PoWiFi is designed to provide high cumulative channel occupancies for power delivery while minimizing the effect on Wi-Fi traffic. To evaluate this, we deploy a PoWiFi router and evaluate its effect on Wi-Fi traffic. We use a Dell Inspiron 1525 laptop with an Atheros chipset as a client associated with our router on channel 1.

We compare four different schemes:

- *Baseline*. PoWiFi is disabled on the router, that is, the router introduces no extra traffic on any of its interfaces.
- *BlindUDP*. The router transmits UDP broadcast traffic at 1 Mbps so as to maximize its channel occupancy.
- *PoWiFi*. The router sends UDP broadcast traffic at 54 Mbps and uses the queue threshold check in Section 3.1.
- *NoQueue*. The router sends UDP broadcast traffic at 54 Mbps but disables the queue threshold check.

We evaluate PoWiFi with various Wi-Fi traffic patterns and metrics: the throughput of UDP and TCP download traffic, the page load time (PLT) of the 10 most popular websites in the United States,<sup>1</sup> and traffic on other Wi-Fi networks in

the vicinity of our benchmarking network.

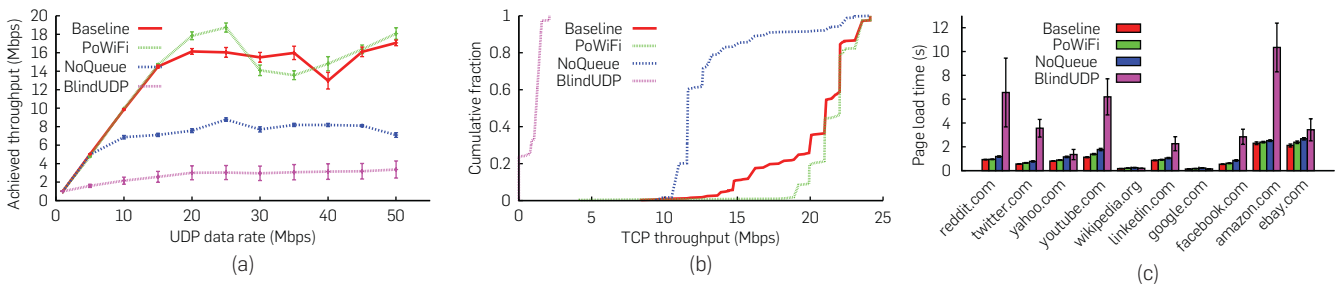
**Effect on UDP traffic.** UDP is a common transport protocol used in media applications such as video streaming. We run `iperf` with UDP traffic to a client 7 ft from the router. The client sets its Wi-Fi bitrate to 54 Mbps and runs five sequential copies of `iperf`, 3 s apart. We repeat the experiments with target UDP data rates between 1 and 50 Mbps, and measure the achieved throughput computed over 500 ms intervals. All the experiments are run during a busy weekday at UW CSE, with multiple other clients and 43 other Wi-Fi networks operating at 2.4 GHz.

Figure 5a plots the average UDP throughput as a function of the 11 tested UDP data rates. The figure shows that *BlindUDP* significantly reduces throughput. With *NoQueue*, the router's kernel does not prioritize the client's `iperf` traffic over the power traffic. This results in roughly a halving of the `iperf` traffic's data rate as the wireless interface is equally shared between the two flows. With PoWiFi, however, the client's `iperf` traffic achieves roughly the same rate as the baseline. This result demonstrates that PoWiFi effectively prioritizes client traffic above its power traffic.

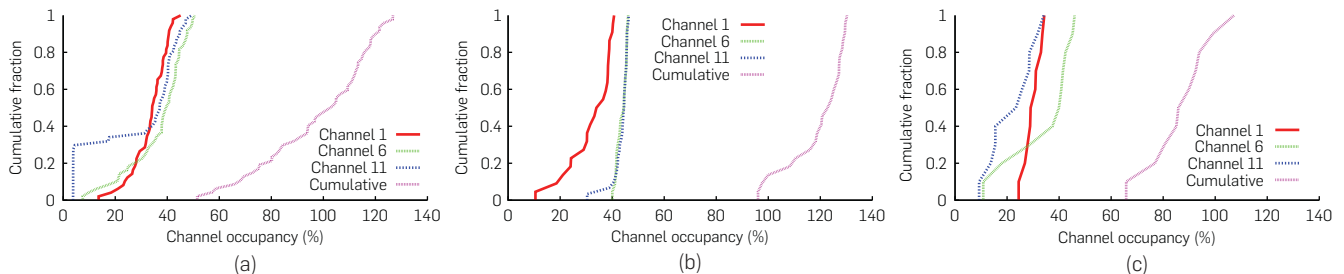
For the PoWiFi experiments above, Figure 6a plots the CDFs of individual channel occupancies on the three Wi-Fi channels. The figure shows that the individual channel occupancies are around 5–50% across the channels. The mean cumulative occupancy, on the other hand, is 97.6%, demonstrating that PoWiFi can efficiently deliver power even in the presence of UDP download traffic.

**Effect on TCP traffic.** Next we run experiments with TCP traffic using `iperf` at the client. The router is configured to run the default Wi-Fi rate-adaptation algorithm. We run experiments over a period of 3 h with a total of 30 runs. In

**Figure 5. Effect on Wi-Fi traffic.** The figures show the effect of various schemes on TCP and UDP throughput as well as the page load times of the top 10 websites in the United States.<sup>1</sup> The plots show that PoWiFi minimizes its effect on the Wi-Fi traffic. (a) UDP experiments, (b) TCP experiments, and (c) PLT experiments.



**Figure 6. PoWiFi channel occupancies.** The plots show the occupancies with PoWiFi for the above UDP (a), TCP (b), and PLT (c) experiments.



each run, we run five sequential copies of iperf, 3 s apart, and compute the achievable throughput over 500 ms intervals, with all the schemes described above.

Figure 5b plots CDFs of the measured throughput values across all the experiments. The plot shows that *BlindUDP* significantly degrades TCP throughput. As before, since *NoQueue* does not prioritize the client traffic over the power packets, it roughly halves the achievable throughput. PoWiFi sometimes achieves higher throughput than the baseline. This is due to changes in channel conditions that occur during the 3-h experiment period. The general trend however points to the conclusion that PoWiFi does not have a noticeable effect on TCP throughput at the client.

Figure 6b plots the CDFs of the channel occupancies for PoWiFi during the above experiments. The figure shows that PoWiFi has a mean cumulative occupancy of 100.9% and hence can efficiently deliver power.

**Effect on PLT.** We develop a test harness that uses the PhantomJS headless browser<sup>11</sup> to download the front pages of the 10 most popular websites in the US<sup>1</sup> 100× each. We clear the cache and pause for 1 s in between page loads. The traffic is recorded with `tcpdump` and analyzed offline to determine PLT and channel occupancy. The router uses the default rate adaptation to modify its Wi-Fi bit rate. The experiments were performed during a busy weekday at UW CSE over a 2-h duration.

Figure 5c shows that *BlindUDP* significantly increases the PLT. This is expected because the 1 Mbps power traffic occupies a much larger fraction of the medium and hence increases packet delays to clients. *NoQueue* improves PLT over *BlindUDP*, with an average delay of 294 ms over the baseline. PoWiFi further minimizes the delay to 101 ms, averaged across websites. This residual delay is due to the computational overhead of PoWiFi from the per-packet checks performed by the kernel. This slows down all the processes in the OS and hence results in additional delays. However, increasing processing power and moving these checks to hardware can further reduce these delays. In our home deployments (Section 6), users did not perceive any noticeable effects on their web performance.

For completeness, we plot the CDFs of channel occupancies for PoWiFi in Figure 6c. The plot shows the same trend as before, with a mean cumulative occupancy of 87.6%.

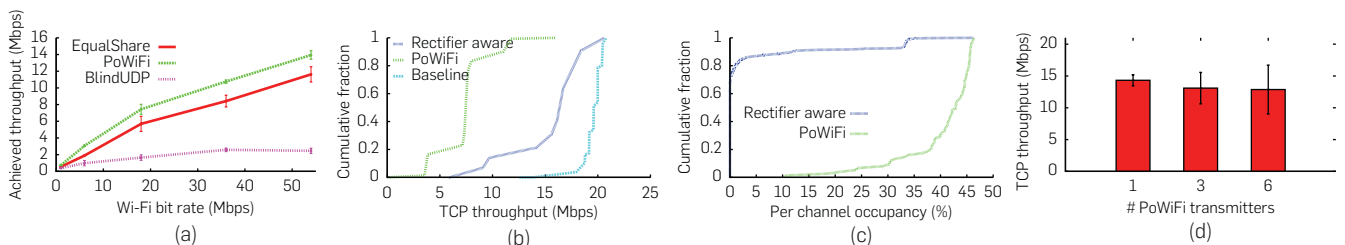
## 4.2. Effect on neighboring Wi-Fi networks

**High cumulative channel occupancy transmissions.** PoWiFi leverages the inherent fairness of the Wi-Fi Medium Access Control to ensure that it is fair to other Wi-Fi networks. As a worst-case evaluation, we consider a situation where PoWiFi always tries to achieve high cumulative channel occupancies at all times. To do this, we place our PoWiFi router in the vicinity of a neighboring Wi-Fi router-client pair operating on channel 1. We configure the PoWiFi router to transmit power packets at the highest achievable channel occupancies using our algorithm on all three nonoverlapping channels. We run iperf with UDP traffic on the neighboring router-client pair at the highest data rate and measure the achievable throughput as before. We repeat the experiments for different Wi-Fi bit rates at the neighboring Wi-Fi router-client pair. We compare three schemes: *BlindUDP* where our router transmits UDP packets at 1 Mbps, *EqualShare* where we set our router to transmit the UDP packets at the same Wi-Fi bit rate as the neighboring router-client pair, and finally PoWiFi. *EqualShare* provides a baseline when every router in the network gets an equal share of the wireless medium.

Figure 7a shows the throughput for the three schemes, averaged across five runs. As expected, *BlindUDP* significantly degrades the neighboring router-client performance. Further, this deterioration is more pronounced at the higher bit rates. With PoWiFi, however, the throughput achieved at the neighboring router-client pair is higher than *EqualShare*. This is because PoWiFi transmits power packets at 54 Mbps; transmissions at such high rates occupy the channel for a smaller duration than, say, a neighboring router transmitting at 16 Mbps. This property means that PoWiFi provides better than equal-share fairness to other Wi-Fi networks. We note that while our experiments are with 802.11g, PoWiFi’s power packets use the highest bit rate available for Wi-Fi. Thus, the above fairness property would hold true even with 802.11n/ac.

**Rectifier-aware power transmissions.** Next, we evaluate the potential of our rectifier-aware technique, to significantly reduce the average channel occupancy of the power transmissions, while efficiently delivering power to the sensors. To do this, we place our battery-free temperature sen-

**Figure 7. Effect of PoWiFi, rectifier aware and concurrent power transmissions on neighboring Wi-Fi networks. The plots show that PoWiFi power transmissions provide better than EqualShare throughput performance. Rectifier aware power transmissions further improve the throughput by reducing the per channel occupancy by a factor of 10. Additionally, increasing the number of concurrently transmitting PoWiFi devices does not degrade the performance of neighboring Wi-Fi devices. (a) PoWiFi bit-rates, (b) Rectifier aware throughput, (c) Rectifier aware occupancies, and (d) Concurrent transmissions.**



sensor close to its maximum operational range at 16 ft from a PoWiFi router; the sensor is set to transmit a temperature value over a UART interface once every minute. The router implements the joint-optimization algorithm from “Rectifier-aware PoWiFi transmissions section.”

We run the experiments for a total of 10 min and observed that the temperature sensor achieves a mean time between updates of 59.93 s with a 0.43 s variance. More importantly, in contrast to transmitting at high channel occupancies (>90%) all the time, our algorithm estimated that the router should transmit for a duration of 9 s with a 80% cumulative occupancy and stay quiet for the remaining time. Figure 7b shows the throughput of an ongoing TCP flow in a neighboring Wi-Fi router–client pair, which shows that the average throughput significantly improves over high-occupancy PoWiFi and is much closer to the baseline throughput without any power packets. Figure 7c shows that rectifier aware transmissions have an average per-channel occupancy of 3.3%, compared to 40% per-channel occupancy for PoWiFi transmissions—a 10× reduction in average occupancy.

**Scalable concurrent power transmissions.** Finally, we provide a proof-of-concept evaluation of our concurrent transmission mechanism. Wi-Fi hardware is designed to turnaround between decoding a packet and transmitting within a Short Interframe Space (SIFS) duration and hence can, in principle, easily achieve the timing requirement in Figure 4d. With only software access to the router, we are limited to implementing PoWiFi timing using high-speed timers and the high-priority queue. Our current software system has 36.15 μs mean turnaround time with 4.61 μs variance.

Using the above mean turnaround time as the silence period, we do a proof-of-concept evaluation. To simplify implementation, we setup a USRP N210 to transmit the pattern in Figure 4 at 30% channel occupancy. The PoWiFi routers join this USRP transmission and concurrently transmit power packets. We evaluate the impact on the TCP throughput of a neighboring Wi-Fi router–client pair as we increase the number of PoWiFi routers. Figure 7d shows that as the number of devices increases, the throughput variance slightly increases. This is because as the number of devices increases, the variance in the turnaround time between Wi-Fi power transmissions increases. The figure, however, shows that the mean throughput is only minimally affected as the number of PoWiFi devices increases from 1 to 6. This shows the feasibility of scaling to multiple PoWiFi routers.

## 5. SENSOR APPLICATIONS

We develop Wi-Fi harvesting sensors at two ends of the energy consumption spectrum: a temperature sensor and a camera. We build both battery-free and battery-recharging versions of each.

### 5.1. Wi-Fi powered temperature sensor

We use our Wi-Fi harvester to convert incoming Wi-Fi signals into DC and power an LMT84 temperature sensor and an MSP430FR5969 microcontroller. The microcontroller reads and transmits sensor data.<sup>14</sup> We optimize our sensor for power and each temperature measurement and transmission operation consumes only 2.77 μJ. In the battery-recharging sensor,

we use two AAA 750 mAh 2.4 V low discharge current NiMH battery and recharge with our battery-charging harvester (see Ref.<sup>14</sup> for more details).

*Experiments.* We evaluate our temperature sensor by measuring the update rate of the sensor as function of operating distance. Specifically, we use a PoWiFi router and place both the battery-recharging and battery-free sensor at increasing distances. In the battery-free case, we measure the update rate by computing the time between successive sensor readings. In the battery-operated case, we measure the battery voltage and the charge current flowing into it from the harvester. Since, each temperature measurement and data transmission takes 2.77 μJ, we compute the ratio of the incoming power to this value to ascertain the sensor update rate for energy-neutral operation. The average cumulative occupancy in our experiments was 91.3%.

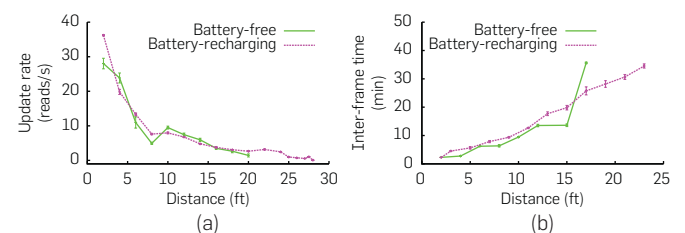
*Results.* Figure 8 shows that the update rate of both battery-recharging and battery-free version of our sensors decrease with distance from the router. This is a result of less power being available and consequently less power being harvested as the distance between the router and sensor increases. Furthermore, we observe that the battery-free sensor operates upto a distance of 20 ft whereas the battery-recharging sensor, optimized for lower input power, has better update rate at distances beyond 15 ft and can operate up to 28 ft from the router.

### 5.2. Wi-Fi powered camera

We use OV7670, a low-power VGA image sensor from Omnivision, interface it with an MSP430FR5969 microcontroller and power it with our harvester. We optimize our firmware for power and achieve a per-image capture energy of 10.4 mJ. On our battery-free camera, we use an ultra-low leakage AVX BestCap 6.8 mF super-capacitor as the storage element. Our battery-recharging camera consists of the same hardware as before, but uses our wirelessly rechargeable 1 mAh lithium-ion coin-cell battery at 3.0 V (see Ref.<sup>14</sup> for details).

*Experiments 1.* We evaluate the camera by measuring the time between successive frames as a function of distance from the router. As before, we use a PoWiFi router—the observed average cumulative occupancy was 90.9% across experiments. At each distance, we wait for the camera to take at least six frames and measure the time between consecutive frames. For the battery-recharging camera, we ascertain the inter-frame duration for an energy-neutral image capture.

**Figure 8. Sensor update rate. The temperature (camera) sensor can operate up to 20 (17) and 28 (23) ft as battery-free and energy-neutral battery-recharging, respectively. (a) Temperature Sensor (b) Camera.**



**Results 1.** Figure 8b shows that the battery-free camera can operate up to 17 ft from the router, with an image capture every 35 min. On the other hand, the battery-recharging camera has an extended range of 23 ft with an image capture every 34.5 min in an energy-neutral manner. Both the sensors have a similar image capture rate up to 15 ft from the router. We also note that Figure 8b limits the range to 23 ft to focus on the smaller values. Our experiments, however, show that the battery-recharging camera can operate up to 26.5 ft with an image capture every 2.6 h.

### 6. HOME DEPLOYMENT STUDY

In Section 4.2, we showed that the channel occupancy of PoWiFi can be optimized for different sensor applications and minimize impact on neighboring Wi-Fi devices. However, PoWiFi's ability to efficiently deliver power depends on the traffic patterns of other Wi-Fi networks in the vicinity as well as the router's own client traffic, both of which can be unpredictable. So, we deploy our system in six homes in a metropolitan area and measure PoWiFi's ability to continuously achieve high channel occupancies.

Table 1 summarizes the number of users, devices, and other 2.4 GHz routers nearby in each of our deployments. We replace the router in each home with a PoWiFi router, and the occupants use it for normal Internet access for 24 h. Our router uses the same SSID and authentication information

**Table 1. Summary of our home deployment**

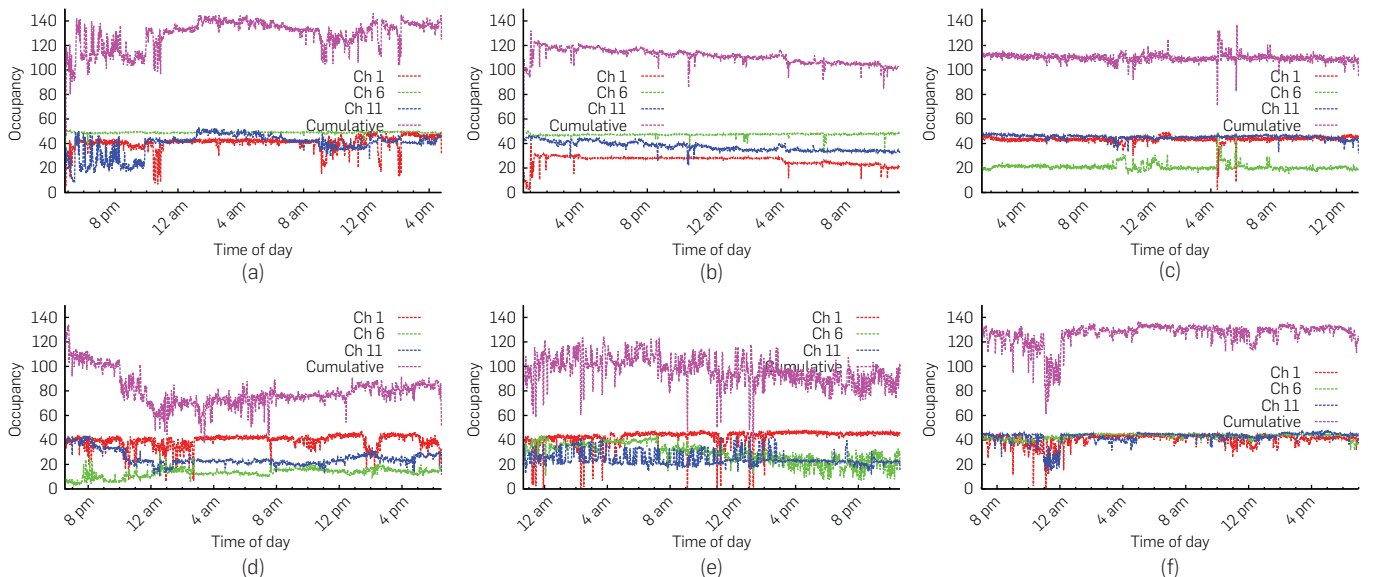
Home #	1	2	3	4	5	6
Users	2	1	3	2	1	3
Devices	6	1	6	4	2	6
Neighboring APs	17	4	10	15	24	16

as the original router, which we disconnect. We place our router within a few feet of the original router, with the exact location determined by user preferences. In all six deployments, we set our router to provide Internet connectivity on channel 1 and to transmit power packets on channels 1, 6, and 11 using the algorithm in Section 3.1.

We log the router's channel occupancy on each of the three Wi-Fi channels at a resolution of 60s. Figure 9 plots the occupancy values for each Wi-Fi channel over the 24-h deployment duration. We also plot the cumulative occupancy across the channels. The figures show that:

- We see significant variation in per-channel occupancy across homes. This is because our router uses carrier sense to enforce fairness with other Wi-Fi networks. It scales back its transmissions on high occupancy channel but when the load on neighboring networks is low, the router occupies a larger fraction of the wireless channel.
- The cumulative occupancy is high over time in all our home deployments. Specifically, the mean cumulative occupancies for the six homes are in the 78–127% range. Although some of these occupancies are much greater than 100%, once can reduce the rate of the power traffic based on the cumulative occupancy to ensure that it is below 100%. PoWiFi does not currently implement this feature.
- The users in homes 1–4 did not perceive any noticeable difference in their user experience. The user in home 5, however, noted a significant improvement in PLTs and better streaming experience. This was primarily because home 5 originally was using a cheap low-grade router with worse specifications. A user in home 6 noted a slight deterioration in YouTube viewing experience for a 30-min duration. Our analysis showed that our router

**Figure 9. PoWiFi channel occupancies in home deployments. We see significant variation in per-channel occupancy values across homes. This is because PoWiFi uses carrier sense that reduces its occupancy when the neighboring networks are loaded. The cumulative occupancy, however, is high across time in all home deployments. (a) Home 1, (b) Home 2, (c) Home 3, (d) Home 4, (e) Home 5, and (f) Home 6.**



occupancy, including both client and power traffic, dipped during this duration. This points to external causes including interference from other devices in the environment.

## 7. RELATED WORK

Early RF power delivery systems were developed as part of RFID systems to harvest small amounts of power from a dedicated 900 MHz UHF RFID readers.<sup>13</sup> The power harvested from RFID signals has been used to operate accelerometers,<sup>13</sup> temperature sensors,<sup>13</sup> and recently cameras.<sup>9</sup> Our efforts on power delivery over Wi-Fi are complimentary to RFID systems. In principle, one can combine multiple ISM bands including 900 MHz, 2.4 GHz, and 5 GHz to design an optimal power delivery system. This paper takes a significant step toward this goal.

Recently, researchers have demonstrated the feasibility of harvesting small amounts of power from ambient TV<sup>7</sup> and cellular base station signals<sup>19</sup> in the environment. While TV and cellular signals are stronger in outdoor environments, they are significantly attenuated indoors limiting the corresponding harvesting opportunities. The ability to power devices using Wi-Fi can augment the above capabilities and enable power harvesting indoors.

Researchers have explored the feasibility of harvesting power in the 2.4GHz ISM band.<sup>4, 10, 15, 18</sup> These efforts have demonstrated power harvesting from continuous wave (CW) transmissions and none have powered devices with existing Wi-Fi chipsets. In contrast, PoWiFi is the first power over Wi-Fi system that works with existing Wi-Fi chipsets and minimizes its impact on Wi-Fi performance. Furthermore, none of the systems power sensors and microcontrollers or recharge batteries and operate at distances demonstrated in this work.

Our work is also related to efforts from startups such as Ossia<sup>2</sup> and Wattup.<sup>21</sup> These claim to deliver around 1W of power at 15 ft and charge a mobile phone.<sup>5</sup> Back-of-the-envelope calculations however show that this requires continuous transmissions with an EIRP (equivalent isotropic radiated power) of 83.3 dBm (213 kW). This not only jams the Wi-Fi channel but also is 50,000× higher power than that allowed by FCC regulations part 15 for point to multi-point links. In contrast, our system is designed to operate within the FCC limits and has minimal impact on Wi-Fi traffic. We note that in the event of an FCC exception to these startups, our multichannel design can be used to deliver such high power without having significant impact on Wi-Fi performance.

## 8. CONCLUSION

There is increasing interest in the Internet of Things where small computing sensors and mobile devices are embedded in everyday objects and environments. A key issue is how to power these devices as they become smaller and more numerous; plugging them in to provide power is inconvenient and is difficult at large scale. We introduce a novel far-field power delivery system that uses existing Wi-Fi chipsets while minimizing the impact on Wi-Fi network performance. While this is a first step toward using Wi-Fi

for power delivery system, we believe that with subsequent iterations of the harvester we can significantly increase the capabilities of our system.

## Acknowledgments

This research is funded in part by NSF grants CNS-1452494 and CNS-1407583, a Qualcomm Innovation Fellowship, a Intel Fellowship, and University of Washington. 

## References

1. Alexa—Top Sites in United States. <http://www.alexa.com/topsites/countries/US>. Loaded January 13, 2015.
2. Cota by Ossia. <http://www.ossiainc.com/>.
3. Covic, G., Boys, J. Inductive power transfer. , 6 (2013), 1276–1289.
4. Curty, J.-P., Joehl, N., Dehollaini, C., Declercq, M.J. Remotely powered addressable uhf rfid integrated system. , 11 (2005), 2193–2202.
5. Energoous Wattup wireless charging demo. <http://www.engadget.com/2015/01/05/energoous-wattup-wireless-charging-demo/>.
6. Gollakota, S., Ahmed, N., Zeldovich, N., Katabi, D. Secure in-band wireless pairing. In (2011). San Francisco, CA, USA, 1–16.
7. Liu, V., Parks, A., Talla, V., Gollakota, S., Wetherall, D., Smith, J.R. Ambient backscatter: Wireless communication out of thin air. , 4 (2013), 39–50.
8. Low, Z.N., Chinga, R., Tseng, R., Lin, J. Design and test of a high-power high-efficiency loosely coupled planar wireless power transfer system. , 5 (2009), 1801–1812.
9. Naderiparizi, S., Parks, A., Kapetanovic, Z., Ransford, B., Smith, J.R. Wispcam: A battery-free rfid camera. In (2015). IEEE, 166–173.
10. Olgun, U., Chen, C.-C., Volakis, J. Design of an efficient ambient WiFi energy harvesting system. , 11 (2012), 1200–1206.
11. PhantomJS. <http://phantomjs.org/>. Loaded January 14, 2015.
12. S-882Z Series by SEIKO. [http://www.eet-china.com/ARTICLES/2006MAY/PDF/S882Z\\_E.pdf](http://www.eet-china.com/ARTICLES/2006MAY/PDF/S882Z_E.pdf).
13. Sample, A., Yeager, D., Powledge, P., Mamishev, A., Smith, J. Design of an rfid-based battery-free programmable sensing platform. , 11 (2008), 2608–2615.
14. Talla, V., Kellogg, B., Ransford, B., Naderiparizi, S., Gollakota, S., Smith, J.R. Powering the next billion devices with Wi-Fi. arXiv preprint arXiv:1505.06815 (2015).
15. Talla, V., Pellerano, S., Xu, H., Ravi, A., Palaskas, Y. Wi-Fi RF energy harvesting for battery-free wearable radio platforms. In (2015). IEEE, 47–54.
16. Tesla, N. . Hart Brothers, Williston, Vermont, 1982.
17. Trotter, M.S., Griffin, J.D., Durgin, G.D. Power-optimized waveforms for improving the range and reliability of rfid systems. In (2009). IEEE, 80–87.
18. Valenta, C., Durgin, G. Harvesting wireless power: Survey of energy-harvester conversion efficiency in far-field, wireless power transfer systems. 4 (2014), 108–120.
19. Visser, H., Reniers, A., Theeuwes, J. Ambient RF energy scavenging: GSM and WLAN power density measurements. In (2008). IEEE, 721–724.
20. Waters, B., Sample, A., Bonde, P., Smith, J. Powering a ventricular assist device (vad) with the free-range resonant electrical energy delivery (free-d) system. , 1 (2012), 138–149.
21. Wattup by Energoous. <http://www.energoous.com/overview/>.

**Vamsi Talla, Bryce Kellogg, Benjamin Ransford, Saman Naderiparizi, Joshua R. Smith, and Shyamnath Gollakota** ([vamsit,samanp,kellogg,jrsjrs,ransford,gshyam@uw.edu](mailto:vamsit,samanp,kellogg,jrsjrs,ransford,gshyam@uw.edu)) Department of Computer Science and Engineering, University of Washington, Seattle, WA.

# Technical Perspective

## Data Distribution for Fast Joins

By Leonid Libkin

WHEN WE TALK about big data and data analytics, a big—some say, the biggest—component of it is what is known as data wrangling: extracting, integrating, querying, and otherwise preparing data for meaningful analytic algorithms to be applied. Data wrangling relies on well-known and trusted database technology, but many classical database questions now are posed in new settings. One reason for this is that parallel processing becomes very important for handling large amounts of data. This has given rise to a steady line of research on classical database problems in new environments where costs caused by massive parallelism dominate the usual I/O costs of the standard database environment. These new costs are primarily related to communication.

What is the most drastic way to reduce the cost of communication for parallel data processing algorithms, for example, query evaluation? If we could distribute data to servers in a single round of communication, let them do their work, and then collect the results to produce the answer to our query, that would be ideal. This is precisely the kind of questions studied in the following paper. It looks at join algorithms: the most common and important task in database query processing, and investigates conditions on joins that make one-round parallel algorithms produce correct results.

They are not the first to look at this problem. In 2010, Afrati and Ullman initiated the study of such multi-join algorithms. A refinement, Hypercube, algorithm was proposed in 2013 by Beame, Koutris, and Suciu. In those algorithms, the network topology is a hypercube. To evaluate a query, one replicates each tuple in several of its nodes and then lets each node perform its computation. While the hypercube is a rather natural distribution policy, it is certainly not the only one. But can we reason about single-round join evaluation under arbitrary distribution policies?

Also, distribution policies are query-dependent. While finding one policy for all scenarios is of course unrealistic, what about a more down-to-earth requirement: if we already know that a policy works for a query  $Q$ , perhaps we can use the same policy for another query  $Q'$ , without redistributing data? This paper addresses these questions.

**The formalism.** It is very simple and elegant. A network is a set of node names; a distribution policy assigns each tuple in a relation to a set of nodes. This is the communication round. The query  $Q$  is then executed locally at each node. It is parallel correct if such a distributed evaluation gives the result of  $Q$ ; that is, tuples in the answer to  $Q$  are exactly those produced locally at network nodes.

Next, if we have two queries  $Q$  and  $Q'$ , and we know that each distribution policy that makes  $Q$  parallel-correct does the same for  $Q'$ , we say that parallel-correctness transfers from  $Q$  to  $Q'$ . In this case, the work done for  $Q$  in terms of looking for the right distribution policy need not be redone for  $Q'$ .

**The results, and what they tell us.** This is a theory paper; the main results are about the complexity of checking parallel-correctness and parallel-transferability. It concentrates on the class of conjunctive queries, that is, slightly more general queries than multi-way joins.

**The following paper looks at join algorithms: the most common and important task in database query processing.**


Parallel-correctness, under mild assumptions, is  $\Pi_2^P$ -complete. That is, it is a bit harder than NP or coNP, but still well within polynomial space. In practice, this means that checking whether a distribution policy guarantees correctness for all databases can be done in exponential time. Note that this is a static analysis problem (the database is not an input), and exponential time is tolerable and in fact the expected best case for conjunctive queries (as their containment is NP-complete).

The authors then show the same problems for conjunctive queries with negations requires (modulo some complexity theory assumptions) *double*-exponential time, that is, is realistically unsolvable, which means one needs to restrict attention to simple joins.

Finally, transferability of parallel-correctness for conjunctive queries is solvable in exponential time (remember, this is a problem about queries, not about data), and importantly it is in NP for many classes of conjunctive queries, like multi-joins (which hints at the possibility of using efficient NP solvers to address this problem in practice).

To conclude, I would like to explain why I view this as a model database theory paper. Such a paper ought to have several key ingredients:

- ▶ It should consider a real data management problem of interest in practice;
- ▶ It should provide a clean and simple formalism that can be followed by theoreticians and practitioners alike;
- ▶ It should provide theoretical results that give us insights about the original practical problem.

The paper ticks all these boxes: It provides an elegant theoretical investigation of a practically important problem, and gives a good set of results that delineate the feasibility boundary. 

**Leonid Libkin** (libkin@inf.ed.ac.uk) is a professor in the School of Informatics and chair of Foundations of Data Management at the University of Edinburgh, Scotland.

Copyright held by author.

# Reasoning on Data Partitioning for Single-Round Multi-Join Evaluation in Massively Parallel Systems

By Tom J. Ameloot\*, Gaetano Geck, Bas Ketsman†, Frank Neven, and Thomas Schwentick

## Abstract

Evaluating queries over massive amounts of data is a major challenge in the big data era. Modern massively parallel systems, such as, Spark, organize query answering as a sequence of rounds each consisting of a distinct communication phase followed by a computation phase. The communication phase redistributes data over the available servers, while in the subsequent computation phase each server performs the actual computation on its local data. There is a growing interest in single-round algorithms for evaluating multiway joins where data is first reshuffled over the servers and then evaluated in a parallel but communication-free way. As the amount of communication induced by a reshuffling of the data is a dominating cost in such systems, we introduce a framework for reasoning about data partitioning to detect when we can avoid the data reshuffling step. Specifically, we formalize the decision problems parallel-correctness and transfer of parallel-correctness, provide semantical characterizations, and obtain tight complexity bounds.

## 1. INTRODUCTION

The background scenario for this work is that of large-scale data analytics where massive parallelism is utilized to answer complex join queries over multiple database tables. For instance, as described by Chu et al.,<sup>7</sup> data analytics engines face new kinds of workloads, where multiple large tables are joined, or where the query graph has cycles. Furthermore, recent in-memory systems (e.g., Refs.<sup>11, 13, 19, 23</sup>) can fit data in main memory by utilizing a multitude of servers. Koutris and Suciu<sup>12</sup> introduced the Massively Parallel Communication (MPC) model to facilitate an understanding of the complexity of query processing on shared-nothing parallel architectures. For such systems, performance is no longer dominated by the number of I/O requests to external memory as in traditional systems but by the communication cost for reshuffling data during query execution. When queries need to be evaluated in several rounds, such reshuffling can repartition the whole database and can thus be very expensive.

While in traditional distributed query evaluation, multi-join queries are computed in several stages over a join tree

possibly transferring data over the network at each step, we focus on query evaluation algorithms within the MPC model that only require *one* round of communication. Such algorithms consist of two phases: a *distribution phase* (where data is repartitioned or reshuffled over the servers) followed by a *computation phase*, where each server contributes to the query answer in isolation, by evaluating the query at hand over the local data without any further communication. We refer to such algorithms as *generic one-round algorithms*. Afrati and Ullman<sup>1</sup> describe an algorithm that computes a multi-join query in a *single* communication round. The algorithm uses a technique that can be traced back to Ganguly et al.<sup>9</sup> Beame et al.<sup>4, 5</sup> refined the algorithm, named it *HyperCube*, and showed that it is a communication-optimal algorithm for single-round distributed evaluation of conjunctive queries.

The generic one-round HyperCube algorithm requires a reshuffling of the base data for *every* separate query. As the amount of communication induced by a reshuffling of the data can be huge, it is important to detect when the reshuffle step can be avoided. We present a framework for reasoning about data partitioning for generic one-round algorithms for the evaluation of queries under *arbitrary* distribution policies, not just those resulting from the HyperCube algorithm. To target the widest possible range of repartitioning strategies, the initial distribution phase is therefore modeled by a distribution policy that can be *any* mapping from facts to subsets of servers.

The optimization framework is motivated by two concrete scenarios. In the first scenario, we assume that the data is already partitioned over the servers and we want to know whether a given query can be evaluated correctly over the given data distribution *without reshuffling the data*. In the second scenario, the data distribution might be unknown or hidden, but it is known that it allowed the correct evaluation of the *previous* query. Here, we ask whether this knowledge

The original version of this paper is entitled “Parallel-Correctness and Transferability for Conjunctive Queries” and was first published in the *Proceedings of the 2015 ACM Symposium on Principles of Database Systems*. A modified version entitled “Data Partitioning for Single-Round Multi-Join Evaluation in Massively Parallel Systems” appeared in the March 2016 issue of *ACM Sigmod Record*.

\* Postdoctoral Fellow of the Research Foundation – Flanders (FWO).

† PhD Fellow of the Research Foundation – Flanders (FWO).

guarantees that the given (next) query can be evaluated correctly without reshuffling. To this end, we formalize the following decision problems:

**Parallel-Correctness:** Given a distribution policy and a query, can we be sure that the corresponding generic one-round algorithm will always compute the query result correctly—no matter the actual data?

**Parallel-Correctness Transfer:** Given two queries  $Q$  and  $Q'$ , can we infer from the fact that  $Q$  is computed correctly under the current distribution policy, that  $Q'$  is computed correctly as well?

We say that parallel-correctness *transfers* from  $Q$  to  $Q'$ , denoted  $Q \xrightarrow{\text{pc}} Q'$ , when  $Q'$  is parallel-correct under every distribution policy for which  $Q$  is parallel-correct. Parallel-correctness transfer is particularly relevant in a setting of automatic data partitioning where an optimizer tries to automatically partition the data across multiple nodes to achieve overall optimal performance for a specific workload of queries (see, e.g., Refs.<sup>15,18</sup>). Indeed, when parallel-correctness transfers from a query  $Q$  to a set of queries  $S$ , then any distribution policy under which  $Q$  is parallel-correct can be picked to evaluate all queries in  $S$  without reshuffling the data.

We focus in this paper on conjunctive queries and first study the parallel-correctness problem. We give a characterization of parallel-correctness: a distribution policy is parallel-correct for a query, if and only if for every *minimal* valuation of the query there is a node in the network to which the distribution assigns all facts required by that valuation. This criterion immediately yields<sup>a</sup> a  $\Pi_2^p$  upper bound for parallel-correctness, for various representations of distribution policies. It turns out that this is essentially optimal, because the problem is actually  $\Pi_2^p$ -complete. These results also hold in the presence of union and inequalities. When negation is added, deciding parallel-correctness might involve counterexample databases of exponential size. More specifically, in the presence of negation deciding parallel-correctness is coNEXPTIME-complete. The latter result is related to the new result that query containment for conjunctive queries with negation is coNEXPTIME-complete, as well.

For parallel-correctness transfer we also first provide a semantical characterization in terms of a (value-based) containment condition for minimal valuations of  $Q'$  and  $Q$  (Proposition 6.4). Deciding transferability of parallel-correctness for conjunctive queries is  $\Pi_3^p$ -complete, again even in the presence of unions and inequalities. We emphasize that the implied exponential time algorithm for parallel-correctness transfer does not rule out practical applicability because the running time is exponential in the size of the queries and not in the size of a database.

<sup>a</sup>In this article, we refer to standard complexity classes like the famous class NP, two classes from the second and third level of the *polynomial hierarchy*,  $\Pi_2^p$  and  $\Pi_3^p$ , respectively, and the exponential time analogon of coNP, coNEXPTIME. More information can be found in any textbook on computational complexity, for example, see Ref.<sup>4</sup>

**Outline.** In Section 2, we introduce the necessary preliminaries regarding databases and conjunctive queries. In Section 3, we discuss the MPC model. In Section 4, we exemplify the HyperCube algorithm. In Sections 5 and 6, we explore parallel-correctness and parallel-correctness transfer. We present concluding remarks together with direction for further research in Section 7.

## 2. CONJUNCTIVE QUERIES

In this article, a (*database*) *instance*  $I$  is a finite set of facts of the form  $R(a_1, \dots, a_n)$ , where  $R$  is an  $n$ -ary relation symbol from a given database schema and each  $a_i$  is an element from some given infinite domain **dom**.

A *conjunctive query* (CQ)  $Q$  is an expression of the form

$$H(\mathbf{x}) \leftarrow R_1(\mathbf{y}_1), \dots, R_m(\mathbf{y}_m),$$

where every  $R_i$  is a relation name, every tuple  $\mathbf{y}_i$  matches the arity of  $R_i$ , and every variable in  $\mathbf{x}$  occurs in some  $\mathbf{y}_i$ . We refer to the *head atom*  $H(\mathbf{x})$  by  $head_Q$  and to the set  $\{R_1(\mathbf{y}_1), \dots, R_m(\mathbf{y}_m)\}$  by  $body_Q$ . We denote by  $vars(Q)$  the set of all variables occurring in  $Q$ .

A *valuation* for a CQ  $Q$  maps its variables to values, that is, it is a function  $V: vars(Q) \rightarrow \mathbf{dom}$ . We refer to  $V(body_Q)$  as the facts *required* by  $V$ . A valuation  $V$  *satisfies*  $Q$  on instance  $I$  if all facts required by  $V$  are in  $I$ . In that case,  $V$  *derives* the fact  $V(head_Q)$ . The *result* of  $Q$  on instance  $I$ , denoted  $Q(I)$ , is defined as the set of facts that can be derived by satisfying valuations for  $Q$  on  $I$ . We denote the class of all CQs by CQ.

**EXAMPLE 2.1.** Let  $I_e$  be the example database instance

$$\{Like(a, b), Like(b, a), Like(b, c), Dislike(a, a), Dislike(c, a)\},$$

and  $Q_e$  be the example CQ

$$H(x_1, x_3) \leftarrow Like(x_1, x_2), Like(x_2, x_3), Dislike(x_3, x_1).$$

Then  $V_1 = \{x_1 \mapsto a, x_2 \mapsto b, x_3 \mapsto a\}$  and  $V_2 = \{x_1 \mapsto a, x_2 \mapsto b, x_3 \mapsto c\}$  are the only satisfying valuations. Consequently,  $Q_e(I_e) = \{H(a, a), H(a, c)\}$ .

## 3. MPC MODEL

The MPC model was introduced by Koutris and Suciu<sup>12</sup> to study the parallel complexity of conjunctive queries. It is motivated by query processing on big data that is typically performed on a shared-nothing parallel architecture where data is stored on a large number of servers interconnected by a fast network. In the MPC model, computation is performed by  $p$  servers connected by a complete network of private channels. Examples of such systems include Pig,<sup>17</sup> Hive,<sup>20</sup> Dremel,<sup>13</sup> and Spark.<sup>23</sup> The computation proceeds in rounds where each round consists of two distinct phases:

- *Communication Phase:* The servers exchange data by communicating with all other servers.
- *Computation Phase:* Each server performs only local computation (on its local data).

The number of rounds then corresponds to the number of synchronization barriers that an algorithm requires. The input data is initially partitioned among the  $p$  servers and every



server receives  $1/p$ th of the data. There are no assumptions on the particular partitioning scheme. At the end of the execution, the output must be present in the union of the  $p$  servers. As the model focuses primarily on quantifying the amount of communication there is no a priori bound on the computational power of a server. A relevant measure is the *load* at each server, which is the amount of data received by a server during a particular round. Examples of optimization goals are minimizing total load (e.g., Ref.<sup>1</sup>) and minimizing maximum load (e.g., Ref.<sup>12</sup>).

To get a feeling for the model, we next present simple examples of single- and multi-round algorithms in the MPC model for evaluating specific conjunctive queries.

**EXAMPLE 3.1.** (1) Consider the query  $Q_1$

$$H(x, y, z) \leftarrow R(x, y), S(y, z),$$

joining two binary relations  $R$  and  $S$  over a common attribute. Let  $h$  be a hash function mapping every domain value to one of the  $p$  servers. The following single-round algorithm computes  $Q_1$ . In the communication phase, executed by every server on its local data, every tuple  $R(a, b)$  is sent to server  $h(b)$  while every tuple  $S(c, d)$  is sent to server  $h(c)$ . In the computation phase, every server evaluates  $Q_1$  on the received data. The output of the algorithm is the union of the results computed at the computation phase. This strategy is called a repartition join in Ref.<sup>6</sup>

(2) Let  $Q_2$  be the triangle query:

$$H(x, y, z) \leftarrow R(x, y), S(y, z), T(z, x).$$

One way to evaluate  $Q_2$  is by two binary joins leading to a two-round algorithm. We assume two hash functions  $h$  and  $h'$ . In the first round, all tuples  $R(a, b)$  and  $S(c, d)$  are sent to servers  $h(b)$  and  $h(c)$ , respectively. The computation phase computes the join of  $R$  and  $S$  at each server in a relation  $K$ . In the second round, each resulting triple  $K(e, f, g)$  is sent to  $h'(e, g)$ , while each tuple  $T(i, j)$  is sent to  $h'(j, i)$ . Finally,  $K$  and  $T$  are joined at each server.

We note that every MapReduce<sup>8</sup> program can be seen as an algorithm within the MPC model since the map phase and reducer phase readily translate to the communication and computation phase of MPC.

#### 4. HYPERCUBE ALGORITHM

To illustrate the HyperCube algorithm, we show in the following example that the triangle query of Example 3.1(2) can be evaluated by a single-round MPC algorithm.

**EXAMPLE 4.1.** Consider again the triangle query  $Q_2$  of Example 3.1(2):

$$H(x, y, z) \leftarrow R(x, y), S(y, z), T(z, x).$$

Let  $\alpha_x, \alpha_y,$  and  $\alpha_z$  be positive natural numbers such that  $\alpha_x \alpha_y \alpha_z = p$ . Every server can then uniquely be identified by a triple in  $[1, \alpha_x] \times [1, \alpha_y] \times [1, \alpha_z]$ . For  $c \in \{x, y, z\}$ , let  $h_c$  be a hash function mapping each domain value to a number in  $[1, \alpha_c]$ . The algorithm then operates as follows. In the communication phase, every fact

- $R(a, b)$  is sent to every server with coordinate  $(h_x(a), h_y(b), \alpha)$  for every  $\alpha \in [1, \alpha_z]$ ; so,  $R(a, b)$  is sent to the subcube determined by the hash values  $h_x(a)$  and  $h_y(b)$  in the  $x$ - and  $y$ -dimension, respectively, as illustrated in Figure 1a;
- $S(b, c)$  is sent to every server with coordinate  $(\alpha, h_y(b), h_z(c))$  for every  $\alpha \in [1, \alpha_x]$ ; and
- $T(c, a)$  is sent to every server with coordinate  $(h_x(a), \alpha, h_z(c))$  for every  $\alpha \in [1, \alpha_y]$ .

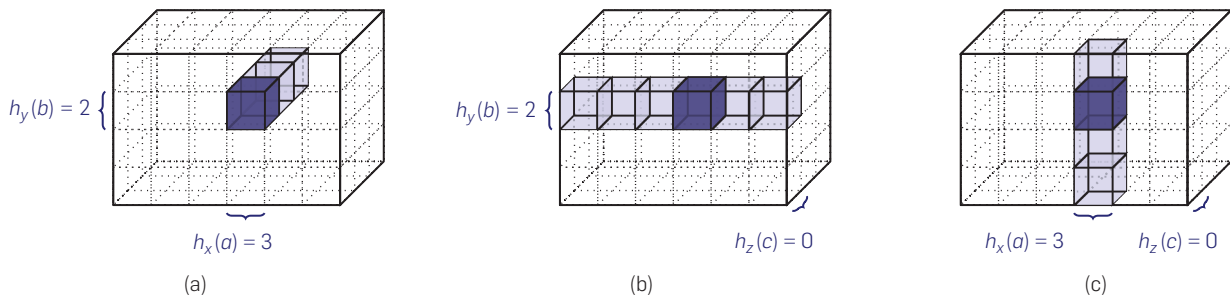
We note that every  $R$ -tuple is replicated  $\alpha_z$  times and similarly for  $S$ - and  $T$ -tuples.

The computation phase consists of evaluating  $Q_2$  on the local data at each server. The algorithm is correct because for every valuation  $V$  for  $Q_2$  some server contains the facts

$$\{V(R(x, y)), V(S(y, z)), V(T(z, x))\},$$

if the (hypothetical) centralized database contains them. In this sense, the algorithm distributes the space of all valuations of  $Q_2$  over the computing servers in an instance independent way through hashing of domain values. In the special case that  $\alpha_x = \alpha_y = \alpha_z = p^{1/3}$ , each tuple is replicated  $p^{1/3}$  times. Assuming each relation consists of  $m$  tuples and there is no skew, each server will receive  $m/p^{2/3}$  tuples for each of the relations  $R, S,$  and  $T$ . So, the maximum load per server is  $O(m/p^{2/3})$ .

**Figure 1. HyperCube distribution policies view the computing nodes in the network as arranged in a multi-dimensional grid. Each dimension corresponds to a variable of the query to be computed. Replication happens in a structurally restricted way: along a line, a plane, or a hyperplane. This figure illustrates the replication of facts  $R(a, b), S(b, c), T(c, a)$  as required by a valuation for the triangle query in Example 4.1 for values  $p = 72, \alpha_x = 6, \alpha_y = 4,$  and  $\alpha_z = 3$ . All facts meet at the node with coordinate  $(h_x(a), h_y(b), h_z(c)) = (3, 2, 0)$ . Therefore the fact  $H(a, b, c)$  can be derived locally, as desired. (a) Replication of  $R(a, b)$ . (b) Replication of  $S(b, c)$ . (c) Replication of  $T(c, a)$ .**



The technique in Example 4.1 can be generalized to arbitrary conjunctive queries and was first introduced in the context of MapReduce by Afrati and Ullman<sup>1</sup> as the *Shares algorithm*. The values  $\alpha_x$ ,  $\alpha_y$ , and  $\alpha_z$  are called shares (hence, the name) and the work of Afrati and Ullman focuses on computing optimal values for the shares minimizing the total load (as a measure for the communication cost).

Beame et al.<sup>4,5</sup> show that the method underlying Example 4.1 is essentially communication optimal for full conjunctive queries  $\mathcal{Q}$ . Assuming that the sizes of all relations are equal to  $m$  and under the assumption that there is no skew, the maximum load per server is bounded by  $O(m/p^{1/\tau^*})$  with high probability. Here,  $\tau^*$  depends on the structure of  $\mathcal{Q}$  and corresponds to the optimal fractional edge packing (which for  $\mathcal{Q}_2$  is  $\tau^* = 3/2$ ). The algorithm is referred to as *HyperCube* in Refs.<sup>4,5</sup> Additionally, the bound is tight over all one-round MPC algorithms, indicating that HyperCube is a fundamental algorithm.

Chu et al.<sup>7</sup> provide an empirical study of HyperCube (in combination with a worst-case optimal algorithm for sequential evaluation<sup>16,22</sup>) for complex join queries, and establish, among other things, that HyperCube performs well for join queries with large intermediate results. However, HyperCube can perform badly on queries with small output.

## 5. PARALLEL-CORRECTNESS

In the remainder of this paper, we present a framework for reasoning about data partitioning for generic one-round algorithms for the evaluation of queries under *arbitrary* distribution policies. We recall from the introduction that such algorithms consist of a distribution phase (where data is repartitioned or reshuffled over the servers) followed by a computation phase where each server evaluates the query at hand over the local data. In particular, generic one-round algorithms are one-round MPC algorithms where every server in the computation phase evaluates the same given query.

When such algorithms are used in a multi-query setting, there is room for optimization. We recall that the HyperCube algorithm requires a reshuffling of the base data for *every* separate query. As the amount of communication induced by a reshuffling of the data can be huge, it is relevant to detect when the reshuffle step can be avoided and the current distribution of the data can be reused to evaluate another query. Here, parallel-correctness and parallel-correctness transfer become relevant static analysis tasks. We study parallel-correctness in this section and parallel-correctness transfer in Section 6.

Before we can address the parallel-correctness problem in detail, we first need to fix our model and our notation.

A characteristic of the HyperCube algorithm is that it reshuffles data on the granularity of facts and assigns each fact in isolation (i.e., independent of the presence or absence of any other facts) to a subset of the servers. This means that the HyperCube reshuffling is independent of the current distribution of the data and can therefore be applied locally at every server. We therefore define distribution policies as arbitrary mappings linking facts to servers.

Following the MPC model, a *network*  $\mathcal{N}$  is a nonempty finite set of node names. A *distribution policy*  $\mathbf{P} = (U, rfacts_p)$  for a

network  $\mathcal{N}$  consists of a universe  $U$  and a total function  $rfacts_p$  that maps each node of  $\mathcal{N}$  to a subset of facts<sup>b</sup> from  $facts(U)$ . Here,  $facts(U)$  denotes the set of all possible facts over  $U$ . A node  $\kappa$  is *responsible for a fact*  $\mathbf{f}$  (under policy  $\mathbf{P}$ ) if  $\mathbf{f} \in rfacts_p(\kappa)$ . For an instance  $I$  and a  $\kappa \in \mathcal{N}$ , let  $loc-inst_{p,I}(\kappa)$  denote  $I \cap rfacts_p(\kappa)$ , that is, the set of facts in  $I$  for which node  $\kappa$  is responsible. We refer to a given database instance  $I$  as the *global instance* and to  $loc-inst_{p,I}(\kappa)$  as the *local instance on node*  $\kappa$ .

The result  $[\mathcal{Q}, \mathbf{P}](I)$  of the distributed evaluation in one round of a query  $\mathcal{Q}$  on an instance  $I$  under a distribution policy  $\mathbf{P}$  is defined as the union of the results of  $\mathcal{Q}$  evaluated over every local instance. Formally,

$$[\mathcal{Q}, \mathbf{P}](I) \stackrel{\text{def}}{=} \bigcup_{\kappa \in \mathcal{N}} \mathcal{Q}(loc-inst_{p,I}(\kappa)).$$

EXAMPLE 5.1. Let  $I_e$  be the example database instance

$$\{Like(a, b), Like(b, a), Like(b, c), Dislike(a, a), Dislike(c, a)\}$$

and  $\mathcal{Q}_e$  be the example CQ

$$H(x_1, x_3) \leftarrow Like(x_1, x_2), Like(x_2, x_3), Dislike(x_3, x_1),$$

from Example 2.1. Consider a network  $\mathcal{N}_e$  consisting of two nodes  $\{\kappa_1, \kappa_2\}$ . Let  $\mathbf{P}_1 = (\{a, b, c\}, rfacts_{p_1})$  be the distribution policy that assigns all *Like*-facts to both nodes  $\kappa_1$  and  $\kappa_2$ , and every fact *Dislike*( $d_1, d_2$ ) to node  $\kappa_1$  when  $d_1 = d_2$  and to node  $\kappa_2$  otherwise. Then,

$$loc-inst_{p_1, I_e}(\kappa_1) = \{Like(a, b), Like(b, a), Like(b, c), Dislike(a, a)\}$$

and

$$loc-inst_{p_1, I_e}(\kappa_2) = \{Like(a, b), Like(b, a), Like(b, c), Dislike(c, a)\}.$$

Furthermore,

$$[\mathcal{Q}_e, \mathbf{P}_1](I_e) = \mathcal{Q}_e(loc-inst_{p_1, I_e}(\kappa_1)) \cup \mathcal{Q}_e(loc-inst_{p_1, I_e}(\kappa_2)),$$

which is just  $\{H(a, b)\} \cup \{H(a, c)\}$ .

We get  $[\mathcal{Q}_e, \mathbf{P}_2](I_e) = \emptyset$  for the distribution policy  $\mathbf{P}_2$  that assigns all *Like*-facts to node  $\kappa_1$  and all *Dislike*-facts to node  $\kappa_2$ .

Now we can define parallel-correctness.

DEFINITION 5.2. A query  $\mathcal{Q}$  is *parallel-correct on instance*  $I$  under distribution policy  $\mathbf{P}$  if  $\mathcal{Q}(I) = [\mathcal{Q}, \mathbf{P}](I)$ .

$\mathcal{Q}$  is *parallel-correct under distribution policy*  $\mathbf{P} = (U, rfacts_p)$ , if it is parallel-correct on all instances  $I \subseteq facts(U)$ .

We note that parallel-correctness is the combination of

- *parallel-soundness*:  $[\mathcal{Q}, \mathbf{P}](I) \subseteq \mathcal{Q}(I)$ , and
- *parallel-completeness*:  $\mathcal{Q}(I) \subseteq [\mathcal{Q}, \mathbf{P}](I)$ .

<sup>b</sup> We mention that for HyperCube distributions, the view is reversed: facts are assigned to nodes. However, both views are essentially equivalent and we will freely adopt the view that fits best for the purpose at hand.

For monotone queries, such as conjunctive queries, parallel-soundness is guaranteed, and therefore parallel-correctness and parallel-completeness coincide.

Whereas Definition 5.2 is in terms of general queries, in the rest of this section, we only consider (extensions of) conjunctive queries.

### 5.1. Conjunctive queries

We first focus on a characterization of parallel-correctness. It is easy to see that a CQ  $Q$  is parallel-correct under distribution policy  $P = (U, rfacts_p)$  if, for each valuation for  $Q$ , the required facts meet at some node. That is, if the following condition holds:

For every valuation  $V$  for  $Q$  over  $U$ , there is a node  $\kappa \in \mathcal{N}$  such that  $V(\text{body}_Q) \subseteq rfacts_p(\kappa)$ . (PC<sub>0</sub>)

However, Condition (PC<sub>0</sub>) is not necessary as the following example shows.

**EXAMPLE 5.3.** Let  $Q_3$  be the CQ

$$H(x, z) \leftarrow R(x, y), R(y, z), R(x, x)$$

and  $V$  the valuation  $\{x \mapsto a, y \mapsto b, z \mapsto a\}$ . Let further  $\mathcal{N} = \{\kappa_1, \kappa_2\}$  and let  $P$  distribute every fact except  $R(a, b)$  onto node  $\kappa_1$  and every fact except  $R(b, a)$  onto node  $\kappa_2$ . Since  $R(a, b)$  and  $R(b, a)$  do not meet under  $P$ , valuation  $V$  witnesses the failure of Condition (PC<sub>0</sub>) for  $P$  and  $Q$ .

However,  $Q_3$  is parallel-correct under  $P$ . Indeed, every valuation that derives a fact  $f$  with the help of the facts  $R(a, b)$  and  $R(b, a)$ , also requires the fact  $R(a, a)$  (or  $R(b, b)$ ). But then,  $R(a, a)$  (or  $R(b, b)$ ) alone is sufficient to derive  $f$  by mapping all variables to  $a$  (or  $b$ ). Therefore, if  $f \in Q(I)$ , for some instance  $I$ , then  $f \in [Q, P](I)$  and thus  $Q_3$  is parallel-correct under  $P$ .

It turns out that it suffices to consider only valuations that are minimal in the following sense.

**DEFINITION 5.4.** A valuation  $V$  for  $Q$  is *minimal* for a CQ  $Q$ , if there is *no* valuation  $V'$  for  $Q$  that derives the same head fact with a strict subset of body facts, that is, such that  $V'(\text{body}_Q) \subsetneq V(\text{body}_Q)$  and  $V(\text{head}_Q) = V'(\text{head}_Q)$ .

**EXAMPLE 5.5.** For a simple example of a minimal valuation and a non-minimal valuation, consider again the CQ  $Q_3$ ,

$$H(x, z) \leftarrow R(x, y), R(y, z), R(x, x).$$

Both valuations  $V_1 = \{x \mapsto a, y \mapsto b, z \mapsto a\}$  and  $V_2 = \{x \mapsto a, y \mapsto a, z \mapsto a\}$  for  $Q_3$  agree on the head variables of  $Q_3$ , but they require different sets of facts. In particular, for  $V_1$  to be satisfying on  $I$ , instance  $I$  must contain the facts  $R(a, b)$ ,  $R(b, a)$ , and  $R(a, a)$ , while  $V_2$  only requires  $R(a, a)$ . Thus  $V_1$  is not minimal for  $Q_3$ . Further, since  $V_2$  requires only one fact it is minimal for  $Q_3$ .

The next proposition shows that it suffices to restrict valuations to minimal valuations in Condition (PC<sub>0</sub>) to get a sufficient *and* necessary condition for parallel-correctness.

**PROPOSITION 5.6.** Let  $Q$  be a CQ. Then  $Q$  is parallel-correct under distribution policy  $P$  if and only if the following holds:

for every minimal valuation  $V$  for  $Q$  over  $U$ , there is a node  $\kappa \in \mathcal{N}$  such that  $V(\text{body}_Q) \subseteq rfacts_p(\kappa)$ . (PC<sub>1</sub>)

We emphasize that the word *minimal* is the only difference between Conditions (PC<sub>0</sub>) and (PC<sub>1</sub>). We now turn to algorithmic questions, that is, we study the following two algorithmic problems, parameterized by classes  $\mathcal{P}$  of distribution policies.

PCI(CQ,  $\mathcal{P}$ )

**Input:**  $Q \in \text{CQ}, P \in \mathcal{P}$ , instance  $I$

**Question:** Is  $Q$  parallel-correct on  $I$  under  $P$ ?

PC(CQ,  $\mathcal{P}$ )

**Input:**  $Q \in \text{CQ}, P \in \mathcal{P}$

**Question:** Is  $Q$  parallel-correct under  $P$ ?

The quantifier structure in Condition (PC<sub>1</sub>) hints at a  $\Pi_2^p$  upper bound for the complexity of parallel-correctness.<sup>c</sup> The exact complexity cannot be judged without having a bound on the number of nodes  $\kappa$  and the complexity of the test  $V(\text{body}_Q) \subseteq rfacts_p(\kappa)$ . The largest classes of distribution policies, for which we established the  $\Pi_2^p$  upper bound, are gathered in the set  $\mathfrak{P}_{\text{npoly}}$ : it contains classes  $\mathcal{P}$  of distribution policies, for which each policy comes with an algorithm  $\mathcal{A}$  and a bound  $n$  on the representation size of nodes in the network, respectively, such that whether a node  $\kappa$  is responsible for a fact  $f$  is decided by  $\mathcal{A}$  *non-deterministically* in time  $\mathcal{O}(n^k)$ , for some  $k$  that depends only on  $\mathcal{P}$ .

It turns out that the problem of testing parallel-correctness is also  $\Pi_2^p$ -hard, even for the simple class  $\mathcal{P}_{\text{fin}}$  of distribution policies, for which all pairs  $(\kappa, f)$  of a node and a fact are explicitly enumerated. Thus, in a sense, Condition (PC<sub>1</sub>) can essentially not be simplified.

**THEOREM 5.7.** Problems PC(CQ,  $\mathcal{P}$ ) and PCI(CQ,  $\mathcal{P}$ ) are  $\Pi_2^p$ -complete, for every policy class  $\mathcal{P} \in \{\mathcal{P}_{\text{fin}}\} \cup \mathfrak{P}_{\text{npoly}}$ .

The upper bounds follow from the characterization in Proposition 5.6 and the fact that pairs  $(\kappa, f)$  can be tested in NP.

We note that Proposition 5.6 continues to hold true in the presence of union and inequalities (under a suitable definition of minimal valuation for unions of CQs) leading to the same complexity bounds as stated in Theorem 5.7.<sup>10</sup>

### 5.2. Conjunctive queries with negation

In this section, we consider conjunctive queries with negation. Specifically, queries can be of the form

$$H(\mathbf{x}) \leftarrow R_1(\mathbf{y}_1), \dots, R_m(\mathbf{y}_m), \neg S_1(\mathbf{z}_1), \dots, \neg S_n(\mathbf{z}_n),$$

where, to ensure safety, we require that every variable in  $\mathbf{x}$  occurs in some  $\mathbf{y}_i$  or  $\mathbf{z}_j$ , and that every variable occurring in a negated atom has to occur in a positive atom as well. A valuation  $V$  now derives a fact  $V(H(\mathbf{x}))$  on an instance  $I$  if every positive atom  $V(R_i(\mathbf{y}_i))$  occurs in  $I$  while none of the negative

<sup>c</sup> Indeed, testing minimality of  $V$  does not introduce another alternation of quantifiers, because it only requires an additional existential quantification of a valuation  $V'$  that serves as a witness, in case  $V$  is not minimal.

atoms  $V(S_j(z_j))$  do. We refer to the class of conjunctive queries with negation as  $CQ^-$ .

We note that, since queries in  $CQ^-$  need not be monotone, parallel-soundness is no longer guaranteed and thus parallel-correctness need not coincide with parallel-completeness.

We illustrate through an example that in the case of conjunctive queries *with negation*, the parallel-correctness problem becomes much more intricate, since it might involve counterexample databases of exponential size. We emphasize that this exponential explosion can only occur if, as in our framework, the arity of the relations in the database schema are not a priori bounded by some constant.

**EXAMPLE 5.8.** Let  $Q_4$  be the following query:

$$H() \leftarrow Val(w_0, w_0), Val(w_1, w_1), \neg Val(w_0, w_1), \\ Val(x_1, x_1), \dots, Val(x_n, x_n), \neg Rel(x_1, \dots, x_n).$$

Let  $P$  be the policy with universe  $U = \{0, 1\}$  and two-node network  $\{\kappa_1, \kappa_2\}$ , which distributes all facts except  $Rel(0, \dots, 0)$  to node  $\kappa_1$  and only fact  $Rel(0, \dots, 0)$  to node  $\kappa_2$ .

Query  $Q_4$  is not parallel-sound under policy  $P$ , due to the counterexample  $I \stackrel{\text{def}}{=} \{Val(0, 0), Val(1, 1)\} \cup \{Rel(a_1, \dots, a_n) \mid (a_1, \dots, a_n) \in \{0, 1\}^n\}$ . Indeed,  $Q_4(I) = \emptyset$  but the all-zero valuation witnesses  $Q_4(loc-inst_{P,I}(\kappa_1)) \neq \emptyset$ .

However,  $I$  has  $2^n + 2$  facts and is a counterexample of minimal size as can easily be shown as follows. First, it is impossible that  $Q_4(I^*) \neq \emptyset$  and  $Q_4(loc-inst_{P,I^*}(\kappa_1)) \neq \emptyset$ , for any  $I^*$ , since  $Rel(0, \dots, 0)$  is the only fact that can be missing at node  $\kappa_1$ , and  $Q_4$  is antimonotonic with respect to  $Rel$ . On the other hand, if  $Q_4(loc-inst_{P,I^*}(\kappa_1)) \neq \emptyset$ , then the literals  $Val(w_0, w_0)$ ,  $Val(w_1, w_1)$ , and  $\neg Val(w_0, w_1)$  ensure that there are at least two different data values (and thus 0 and 1) in  $I^*$ . But then  $Q_4(I^*) = \emptyset$  can only hold if all  $2^n n$ -tuples over  $\{0, 1\}$  are in  $I^*$ .

Although this example requires an exponential size counterexample, in this particular case, the existence of the counterexample is easy to conclude. However, the following result shows that, in general, there is essentially no better algorithm than guessing an exponential size counterexample.

**THEOREM 5.9. (Geck et al.<sup>10</sup>)** For every class  $\mathcal{P} \in \mathfrak{P}_{\text{npoly}}$  of distribution policies, testing parallel completeness for  $UCQ^-$  is coNEXPTIME-complete, and likewise for parallel soundness and correctness.

The result and, in particular, the lower bound even holds if  $\mathfrak{P}_{\text{npoly}}$  is replaced by the class  $\mathfrak{P}_{\text{poly}}$ , where the decision algorithm for pairs  $(\kappa, f)$  is deterministic.

The proof of the lower bounds comes along an unexpected route and exhibits a reduction from query containment for  $CQ^-$  to parallel-correctness for  $CQ^-$ . Query containment asks whether for two queries  $Q$  and  $Q'$ , it holds that  $Q(I) \subseteq Q'(I)$ , for all instances  $I$ . It is shown in Ref.<sup>10</sup> that query containment for  $CQ^-$  is coNEXPTIME-complete, implying coNEXPTIME-hardness for parallel-correctness as well. The result regarding containment of  $CQ^-$  confirms the observation in Ref.<sup>14</sup> that the  $\Pi_2^p$ -completeness result for query containment for  $CQ^-$  mentioned in Ref.<sup>21</sup> only holds for fixed database schemas (or a fixed arity bound, for that matter).

## 6. PARALLEL-CORRECTNESS TRANSFER

Parallel-correctness is relative to a distribution policy. The idea of parallel-correctness *transfer* is to drop this dependence and to infer that a distribution policy is parallel-correct for the next query from the fact that it is parallel-correct for the current query.

**DEFINITION 6.1.** For two queries  $Q$  and  $Q'$ , *parallel-correctness transfers from  $Q$  to  $Q'$*  if  $Q'$  is parallel-correct under every distribution policy for which  $Q$  is parallel-correct. In this case, we write  $Q \xrightarrow{\text{pc}} Q'$ .

**EXAMPLE 6.2.** We consider a database of document IDs with a reference relation  $R$  among them: fact  $R(22, 44)$  states that document 22 references document 44. Query  $Q: H(d_1, d_2) \leftarrow R(d_1, d_2), R(d_1, d_3), R(d_3, d_2)$  asks for all documents  $d_1, d_2$  such that  $d_1$  references  $d_2$  directly as well as in two steps, that is, hopping over a document  $d_3$ .

One might expect that the syntactic subquery

$$Q': H(d_1, d_2) \leftarrow R(d_1, d_3), R(d_3, d_2)$$

which asks for a two-step reference only, is parallel-correct under every distribution policy that allows correct evaluation of query  $Q$ . However, this is not the case because for derived facts  $(i, i)$ , where document  $i$  references *itself* directly and in two steps (taking  $d_3$  as  $i$  as well), query  $Q'$  requires both facts  $R(i, j)$  and  $R(j, i)$  to be present at some node for some  $j$ , while  $Q$  requires only  $R(i, i)$  to be present. See Figure 2 for an example instance and distribution.

Parallel-correctness does transfer from a similar query,  $Q'': H(d_1, d_2, d_3) \leftarrow R(d_1, d_2), R(d_1, d_3), R(d_3, d_1)$ , where  $d_3$  is part of the head, to  $Q'$  because all valuations for  $Q''$  are minimal and every valuation for  $Q'$  requires a subset of the facts required by the same valuation for  $Q''$ .

Like for parallel-correctness, the characterization of parallel-correctness transfer is in terms of minimal valuations. It turns out that the following notion yields the desired semantical characterization.

**DEFINITION 6.3.** For two CQs  $Q$  and  $Q'$ , we say that  $Q$  *covers*  $Q'$  if the following holds:

for every minimal valuation  $V'$  for  $Q'$ , there is a minimal valuation  $V$  for  $Q$ , such that  $V'(body_{Q'}) \subseteq V(body_Q)$ .

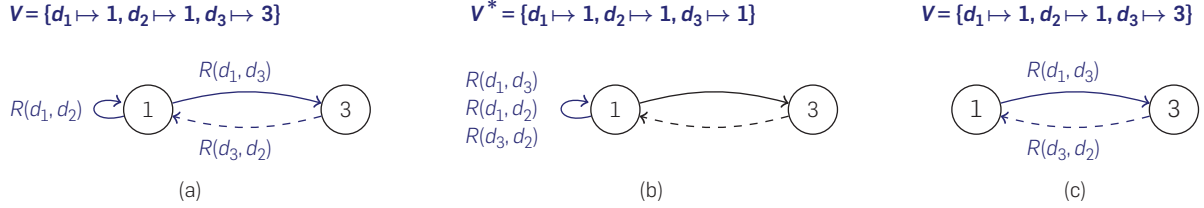
**PROPOSITION 6.4.** For two CQs  $Q$  and  $Q'$ , *parallel-correctness transfers from  $Q$  to  $Q'$  if and only if  $Q$  covers  $Q'$* .

One might be tempted to assume that parallel-correctness transfer is somehow directly linked with query containment. However, as the following example shows, this is not the case.

**EXAMPLE 6.5.** We consider the following four queries:

$$Q_1 : H() \leftarrow S(x), R(x, x), T(x). \\ Q_2 : H() \leftarrow R(x, x), T(x). \\ Q_3 : H() \leftarrow S(x), R(x, y), T(y). \\ Q_4 : H() \leftarrow R(x, y), T(y).$$

**Figure 2.** Global instances  $I = \{R(1, 1), R(1, 3), R(3, 1)\}$  (left and middle) and  $I' = \{R(1, 3), R(3, 1)\}$  (right) are represented by graphs. Solid edges (facts) are located at node  $\kappa_1$ , dashed edges at node  $\kappa_2$ ; colored edges are required by the valuation under concern. Instance  $I$  has globally and locally satisfying valuations for query  $Q$ , subinstance  $I' \subseteq I$  has a globally satisfying valuation but no locally satisfying one—under the same distribution policy. There is thus a policy under which  $Q$  is parallel-correct but  $Q'$  is not, and therefore parallel-correctness does not transfer from  $Q$  to  $Q'$ . (a) Valuation  $V$  satisfies  $Q$  globally on  $I$  but not locally. It is not minimal for  $Q$ . (b) Valuation  $V^*$  satisfies  $Q$  globally on  $I$  and locally on  $\kappa_1$ . It is minimal for  $Q$  and derives the same fact as  $V$ . (c) Valuation  $V$  satisfies  $Q'$  globally on  $I'$  but not locally. It is minimal for  $Q'$ . It does not satisfy  $Q$ .



**Figure 3.** Relationship between the queries of Example 6.5 with respect to (a) parallel-correctness transfer (pc) and (b) query containment ( $\subseteq$ ).

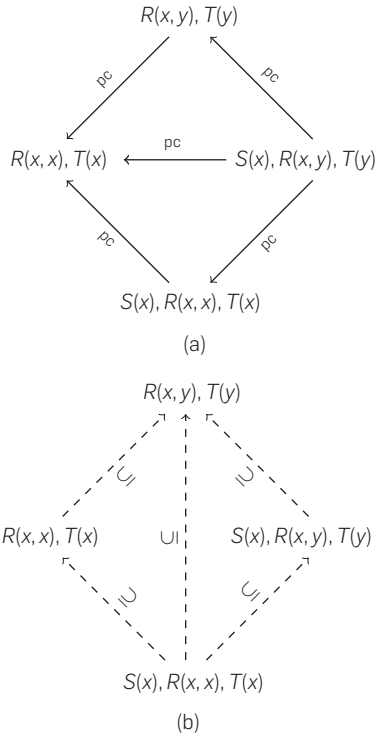


Figure 3a shows how these queries relate with respect to parallel-correctness transfer. As an example,  $Q_3 \xrightarrow{pc} Q_1$ . Figure 3b illustrates that these relationships are completely orthogonal to query containment. Indeed, there are examples where parallel-correctness transfer and query containment coincide ( $Q_3$  vs.  $Q_4$ ), where they hold in opposite directions ( $Q_4$  vs.  $Q_2$ ) and where one but not the other holds ( $Q_3$  vs.  $Q_2$  and  $Q_1$  vs.  $Q_4$ , respectively).

Proposition 6.4 allows us to pinpoint the complexity of parallel-correctness transferability. For a formal statement we define the following algorithmic problem:

**PC-TRANS (CQ)**

**Input:** Queries  $Q$  and  $Q'$  from CQ

**Question:** Does parallel-correctness transfer from  $Q$  to  $Q'$ ?

When the defining condition of “covers” is spelled out by rewriting “minimal valuations” one gets a characterization with a  $\Pi_3$ -structure. Again, it can be shown that this is essentially optimal.

**THEOREM 6.6.** *Problem PC-TRANS(CQ) is  $\Pi_3^P$ -complete.*

The upper bounds follow directly from the characterization in Proposition 6.4. We note that the same complexity bounds continue to hold in the presence of inequalities and for unions of conjunctive queries.<sup>2</sup>

The complexity of transferability is considerably better for a restricted class of conjunctive queries that we call strongly minimal.

**DEFINITION 6.7.** A CQ query is *strongly minimal* if all its valuations are minimal.

Strong minimality generalizes two particularly simple classes of queries:

**LEMMA 6.8.** *A CQ  $Q$  is strongly minimal*

- if it is a full query;
- if it contains no self-joins (every relation name occurs at most once).

**THEOREM 6.9.** *PC-TRANS(CQ) restricted to inputs  $(Q, Q')$ , where  $Q$  is strongly minimal, is NP-complete.*

## 7. CONCLUSION

Parallel-correctness serves as a framework for studying correctness and implications of data partitioning in the context of one-round query evaluation algorithms. A main insight of the work up to now is that testing for parallel-correctness as

well as the related problem of parallel-correctness transfer boils down to reasoning about minimal valuations (of polynomial size) in the context of conjunctive queries (even in the presence of union and inequalities) but seems to require to reason about databases of exponential size when negation is allowed.

There are many questions left unexplored and we therefore highlight possible directions for further research.

From a foundational perspective, it would be interesting to explore the decidability boundary for parallel-correctness and transfer when considering more expressive query languages or even other data models. Obviously, the problems become undecidable when considering first-order logic, but one could consider monotone languages or, for instance, guarded fragment queries. At the same time, it would be interesting to find settings that render the problems tractable, for instance, by restricting the class of queries or by limiting to certain classes of distribution policies.

Parallel-correctness transfer is a rather strong notion as it requires that a query  $Q'$  is parallel-correct for every distribution policy for which another query  $Q$  is parallel-correct. From a practical perspective, however, it could be interesting to determine, given  $Q$  and  $Q'$ , whether there is at least one distribution policy under which both queries are correct. Other questions concern the least costly way to migrate from one distribution to another. As an example, assume a distribution  $P$  on which  $Q$  is parallel-correct but  $Q'$  is not. Find a distribution  $P'$  under which  $Q'$  is parallel-correct and that minimizes the cost to migrate from  $P$  to  $P'$ . Similar questions can be considered for a workload of queries.

Even though the naive one-round evaluation model considered in this paper suffices for HyperCube, it is rather restrictive. Other possibilities are to consider more complex aggregator functions than union and to allow for different queries than the original one to be executed at computing nodes. Furthermore, it could be interesting to generalize the framework beyond one-round algorithms, that is, toward evaluation algorithms that comprise of several rounds. □

## References

- Afrati, F.N., Ullman, J.D. Optimizing multiway joins in a map-reduce environment. *IEEE Trans. Knowl. Data Eng.* 23, 9 (2011), 1282–1298.
- Ameloot, T.J., Geck, G., Ketsman, B., Neven, F., Schwentick, T. Parallel-correctness and transferability for conjunctive queries, submitted for journal publication (2015).
- Arora, S., Barak, B. *Computational Complexity – A Modern Approach*. Cambridge University Press, 2009.
- Beame, P., Koutris, P., Suciu, D. Communication steps for parallel query processing. In *Proceedings of the 32nd Symposium on Principles of Database Systems, PODS'13* (2013), 273–284.
- Beame, P., Koutris, P., Suciu, D. Skew in parallel query processing. In *Proceedings of the 33rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS'14* (2014), 212–223.
- Blanas, S., Patel, J.M., Ercegovac, V., Rao, J., Shekita, E.J., Tian, Y. A comparison of join algorithms for log processing in mapreduce. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010*, A.K. Elmagarmid and D. Agrawal, eds. (Indianapolis, Indiana, USA, June 6–10, 2010). ACM 975–986.
- Chu, S., Balazinska, M., Suciu, D. From theory to practice: Efficient join query evaluation in a parallel database system. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (2015), 63–78.
- Dean, J., Ghemawat, S. MapReduce: Simplified data processing on large clusters. *Commun. ACM* 51, 1 (2008), 107–113.
- Ganguly, S., Silberschatz, A., Tsur, S. Parallel bottom-up processing of datalog queries. *J. Log. Program.* 14, 1&2 (1992), 101–126.
- Geck, G., Ketsman, B., Neven, F., Schwentick, T. Parallel-correctness and containment for conjunctive queries with union and negation. In *International Conference on Database Theory* (2016), 9:1–9:17.
- Halperin, D., Teixeira de Almeida, V., Choo, L.L., Chu, S., Koutris, P., Moritz, D., Ortiz, J., Ruamviboonsuk, V., Wang, J., Whitaker, A., Xu, S., Balazinska, M., Howe, B., Suciu, D. Demonstration of the Myria big data management service. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD'14* (2014), 881–884.
- Koutris, P., Suciu, D. Parallel evaluation of conjunctive queries. In *Proceedings of the 30th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2011*, M. Lenzerini and T. Schwentick, eds. (Athens, Greece, June 12–16, 2011). ACM, 223–234.
- Melnik, S., Gubarev, A., Long, J.J., Romer, G., Shivakumar, S., Tolton, M., Vassilakis, T. Dremel: Interactive analysis of web-scale datasets. *Proc. VLDB Endow.* 3, 1–2 (Sept. 2010), 330–339.
- Mugnier, M., Simonet, G., Thomazo, M. On the complexity of entailment in existential conjunctive first-order logic with atomic negation. *Inf. Comput.* 215 (2012), 8–31.
- Nehme, R., Bruno, N. Automated partitioning design in parallel database systems. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, SIGMOD'11* (2011), 1137–1148.
- Ngo, H.Q., Porat, E., Ré, C., Rudra, A. Worst-case optimal join algorithms. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2012* (2012), 37–48.
- Olston, C., Reed, B., Srivastava, U., Kumar, R., Tomkins, A. Pig latin: A not-so-foreign language for data processing. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008*, J. Tsong and L. Wang, eds. (Vancouver, BC, Canada, June 10–12, 2008). ACM 1099–1110.
- Rao, J., Zhang, C., Megiddo, N., Lohman, G. Automating physical database design in a parallel database. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, SIGMOD'02* (2002), 558–569.
- Shute, J., Vingralek, R., Samwel, B., Handy, B., Whipkey, C., Rollins, E., Oancea, M., Littlefield, K., Menestrina, D., Ellner, S., Cieslewicz, J., Rae, I., Stancescu, T., Apte, H. F1: A distributed sql database that scales. *Proc. VLDB Endow.* 6, 11 (Aug. 2013), 1068–1079.
- Thusoo, A., Sarma, J.S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wyckoff, P., Murthy, R. Hive: A warehousing solution over a map-reduce framework. *VLDB* 2, 2 (2009), 1626–1629.
- Ullman, J.D. Information integration using logical views. *Theor. Comput. Sci.* 239, 2 (2000), 189–210.
- Veldhuizen, T.L. Triejoin: A simple, worst-case optimal join algorithm. In *Proceedings of the 17th International Conference on Database Theory (ICDT)* (2014), 96–106.
- Xin, R.S., Rosen, J., Zaharia, M., Franklin, M.J., Shenker, S., Stoica, I. Shark: Sql and rich analytics at scale. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, SIGMOD'13* (2013), 13–24.

Tom J. Ameloot, Bas Ketsman, and Frank Neven ({tom.ameloot, bas.ketsman, frank.neven}@uhasselt.be), Hasselt University and Transnational University of Limburg, Hasselt, Belgium.

Gaetano Geck and Thomas Schwentick ({gaetano.geck, thomas.schwentick}@udo.edu), TU Dortmund University, Dortmund, Germany.

Copyright held by owner/author.  
Publication rights licensed to ACM. \$15.00.

## **Purdue University** **Department of Computer Science** **Continuing Lecturer**

The Department of Computer Science at Purdue University invites applications for a Continuing Lecturer position beginning August 2017. This position is a non-tenure track instructor position. Duties include teaching and development of computer science undergraduate lecture and laboratory courses. M.S. degree in Computer Science or related field is required, PhD is preferred, with at least 3 years of teaching experience, have familiarity with computer science undergraduate curriculum, strong familiarity with common programming languages, and be able to teach lower division courses. A successful candidate will have interest in and ability to teach large lecture sections, interact with students in small laboratory sections, and train and supervise a large number of undergraduate teaching assistants. A strong commitment to excellence in teaching and exceptional organizational skills is expected.

This position carries competitive salary and benefits. A continuing lecturer will have access to

world class departmental and university computing facilities in addition to computing equipment for the preparation and delivery of course material. Further information about the department can be found at <http://www.cs.purdue.edu>.

Review of applications will begin on February 1, 2017, and continue until the position is filled. Applicants are strongly encouraged to apply online at <https://hiring.science.purdue.edu> by submitting a curriculum vitae, a statement of teaching interests and objectives, and names and contact information of at least three references.

Alternatively, hardcopy applications can be sent to: Faculty Search Chair, Department of Computer Science, 305 N. University Street, Purdue University, West Lafayette, IN 47907. A background check will be required for employment. Purdue University is an EEO/AA employer. All individuals, including minorities, women, individuals with disabilities, and veterans are encouraged to apply.

### **Requirements:**

M.S. degree in Computer Science or related field is required, PhD is preferred, with at least 3 years of teaching experience, have familiarity with

computer science undergraduate curriculum, strong familiarity with common programming languages, and be able to teach lower division courses.

## **Texas State University** **Department of Computer Science** **Assistant Professor**

Applications are invited for multiple tenure-track Assistant Professor positions in the Department of Computer Science to start the fall 2017 semester. Consult the department's faculty employment page at [www.cs.txstate.edu/employment/faculty/](http://www.cs.txstate.edu/employment/faculty/) for job duties, qualifications, application procedure, and information about the department and the university.

Texas State University, to the extent not in conflict with federal or state law, prohibits discrimination or harassment on the basis of race, color, national origin, age, sex, religion, disability, veteran's status, sexual orientation, gender identity or expression. Texas State University is a member of The Texas State University System. Texas State University is an EOE.

## *The National Academies of* **SCIENCES • ENGINEERING • MEDICINE**

### **ARL Distinguished Postdoctoral Fellowship**

The Army Research Laboratory (ARL) Distinguished Postdoctoral Fellowship provides opportunities to pursue independent research that supports the mission of ARL. The Fellow benefits by having the opportunity to work alongside some of the nation's best scientists and engineers. ARL benefits by the expected transfer of new science and technology that enhances the capabilities of the U.S. Army and the warfighter in times of both peace and war.

We invite extraordinary young researchers to participate in this excitement as ARL Distinguished Postdoctoral Fellows. These Fellows must display exceptional abilities in scientific research, and show clear promise of becoming outstanding future leaders. Candidates are expected to have already tackled successfully a major scientific or engineering problem, or have provided a new approach or insight, as evidenced by a recognized impact in their field. ARL offers four named Fellowships, honoring distinguished researchers and work that has been performed at ARL.

The ARL Distinguished Postdoctoral Fellowships are three-year appointments, beginning on October 1 of each year. The annual stipend is \$100,000, and the award includes benefits and potential additional funding for the chosen proposal. A Ph.D. awarded within the past three years at the time of application is required. For more information and to apply, go to [www.nas.edu/arl](http://www.nas.edu/arl).

**Applications must be received by May 1, 2017**

# ARL

## **Renaissance Technologies,**

a quantitative investment management company trading in global financial markets, has openings for researchers and programmers at our Long Island, New York, research campus.

### **The ideal research candidate will have**

- a PhD in computer science, mathematics, physics, statistics, or a related discipline
- a demonstrated capacity to do first-class scientific research
- computer programming skills

### **The ideal programming candidate will have**

- strong analytical and programming skills
- an in-depth knowledge of software development in a C++ Unix environment

Experience in finance is not required.

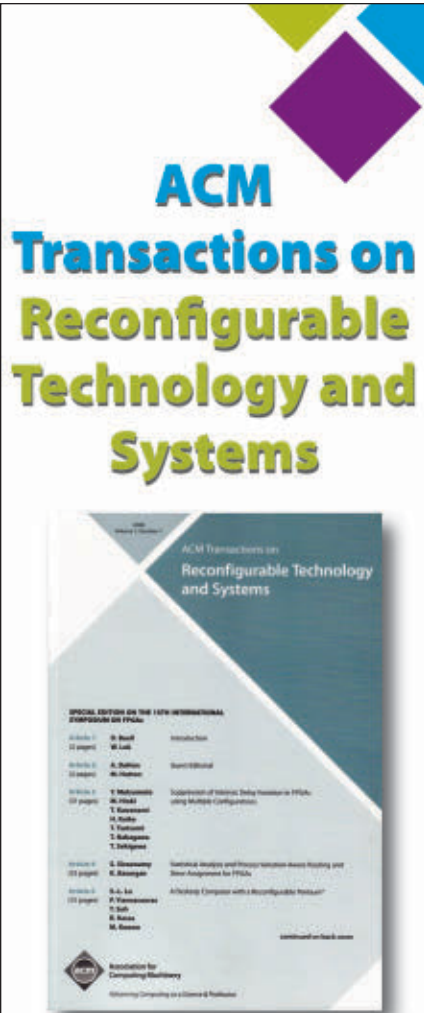
*"Renaissance is . . . the pinnacle  
of quant investing.  
No one else is even close."*

– Bloomberg Markets article,  
November 21, 2016

To apply,  
send your resume  
to [careers@rentec.com](mailto:careers@rentec.com).

For more information, visit  
[www.rentec.com/careers](http://www.rentec.com/careers).

**Renaissance** 



This quarterly publication is a peer-reviewed and archival journal that covers reconfigurable technology, systems, and applications on reconfigurable computers. Topics include all levels of reconfigurable system abstractions and all aspects of reconfigurable technology including platforms, programming environments and application successes.

[www.acm.org/trets](http://www.acm.org/trets)  
[www.acm.org/subscribe](http://www.acm.org/subscribe)



**The University of Alabama in Huntsville  
 Assistant Professor**

The Department of Computer Science of The University of Alabama in Huntsville (UAH) invites applicants for a tenure-track faculty position at the Assistant Professor level beginning August 2017. The incumbent will augment the department's emphases in at least one of the following areas: cloud computing, particularly secure cloud computing; mobile computing, particularly secure mobile computing; or data science, particularly big data applications. Outstanding candidates who couple cybersecurity with other areas of computing could also be considered.

A Ph.D. in computer science or a closely related area is required. The successful candidate will have a strong academic background, perform funded research, be able to carry out research in areas typical for publication in well-regarded academic conference and journal venues, and be keen on undergraduate education.

The department has a strong commitment to excellence in teaching, research, and service; the hire should have good communication, strong teaching potential, and research accomplishments.

UAH is located in an expanding, high technology area, next door to one of the largest research parks in the nation. Nearby are the NASA Marshall Space Flight Center, the Army's Redstone Arsenal, and many high-tech industries. UAH also has an array of research centers, including in information technology, modeling and simulation, etc. In short, collaborative research opportunities are abundant, and many well-educated and highly technically skilled people are in the area. There is also access to excellent public schools and inexpensive housing.

UAH has approximately 8500 students. UAH Computer Science offers the BS, MS, and PhD

degrees in Computer Science and the MS and PhD degrees in modeling and simulation. Approximately 550 undergraduate majors and 175 graduate students are associated with the unit. Faculty research interests are many and include cybersecurity, mobile computing, data science, software engineering, visualization, graphics and game computing, multimedia, AI, image processing, pattern recognition, and distributed systems. Recent NSF figures indicate the department ranks 30th in the nation in overall federal research funding.

**Interested parties should submit a detailed resume with references to [info@cs.uah.edu](mailto:info@cs.uah.edu) or Chair, Search Committee, Dept. of Computer Science The University of Alabama in Huntsville, Huntsville, AL 35899.** Qualified female and minority candidates are encouraged to apply. Initial review of applicants will begin immediately and continue until a suitable candidate is found.

UAH is an equal opportunity/affirmative action institution.

**The University of California,  
 Santa Barbara  
 Faculty Position in Neuroengineering**

The College of Engineering at UCSB invites applications for a faculty position in Neuroengineering with a start date of fall quarter, 2017.

Please visit <https://recruit.ap.ucsb.edu/apply/JPF00950>

The University of California is an Equal Opportunity/Affirmative Action Employer.

All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender identity, national origin, disability status, protected veteran status, or any other characteristic protected by law.



**ADVERTISING IN CAREER OPPORTUNITIES**

**How to Submit a Classified Line Ad: Send an e-mail to [acmm mediasales@acm.org](mailto:acmm mediasales@acm.org). Please include text, and indicate the issue/or issues where the ad will appear, and a contact name and number.**

**Estimates: An insertion order will then be e-mailed back to you. The ad will be typeset according to CACM guidelines. NO PROOFS can be sent. Classified line ads are NOT commissionable.**

**Rates: \$325.00 for six lines of text, 40 characters per line. \$32.50 for each additional line after the first six. The MINIMUM is six lines.**

**Deadlines: 20th of the month/2 months prior to issue date. For latest deadline info, please contact:**

[acmm mediasales@acm.org](mailto:acmm mediasales@acm.org)

**Career Opportunities Online: Classified and recruitment display ads receive a free duplicate listing on our website at:**

<http://jobs.acm.org>

**Ads are listed for a period of 30 days.**

**For More Information Contact:**

**ACM Media Sales  
 at 212-626-0686 or**

[acmm mediasales@acm.org](mailto:acmm mediasales@acm.org)



[CONTINUED FROM P. 104] I was reading various things about where the field was and why it was stuck, and I was exploring different questions. Another key researcher in the area, Johan Håstad from Sweden, was visiting the Institute for Advanced Studies in Princeton, and the interaction with him helped me a lot.

**In fact, you were working on a problem that Håstad proposed when you devised the UGC in 2002.**

Yes, it was about the hardness of approximating 2SAT. Johan had, in some sense, already solved the problem half-way through, and I was thinking about what to do about the second half. Somehow, one fine day, I observed that if one is willing to make the Unique Games Conjecture, then it would solve the second half. It also seemed like a proposal that could break the gridlock in the field. It was a fairly natural proposal to make, but somehow it had not been proposed before, and even after I proposed it, nobody—including me—thought it would be so important.

**Can you describe what the UGC posits?**

It's really about one specific problem, about a system of linear equations over, say, integers, with two variables in each equation, and one seeks an assignment that satisfies the maximum number of equations. We do not know whether this problem is hard to solve or not. The conjecture simply states that yes, it is hard to solve. More specifically, the conjecture states that even if there is an assignment that satisfies 99% of the equations, one cannot efficiently find an assignment that satisfies even 1% of the equations.

**So what are the implications?**

In computer science, the best way to show that a certain problem is hard is to take another problem that is already known to be hard, and then reduce that problem to the first problem. So suppose A and B are two problems, and I already know that A is hard. If I show that A reduces to B, then I can conclude that B must also be hard. Of course, for these reductions to work, I need to start with a hard problem A that I can reduce to other problems, and this is what the UGC does; it identifies a very concrete problem A, as described above, and

**“Broadly speaking, I’m very interested in the interaction between computer science and mathematics, especially geometry and analysis as mathematical areas.”**

conjectures that A is hard. Then, if you believe that A is hard, you can reduce A to many, many other problems that researchers have been very interested in, and prove that those are hard, too.

**So in a single shot, you prove that a wide range of problems that researchers have been interested in are hard.**

Yes, that's correct. When I first described the proposal to Sanjeev and Johan back in 2002, they were kind of lukewarm about it. Even to me, its full significance wasn't clear. But I still felt it was worth writing up and publishing a paper about it. And then in the next 10 to 15 years, the consequences started emerging, and there was a slow realization about how important this problem is as a starting point.

**Other research directions emerged from the UGC, as well.**

Yes, to begin with, one can investigate whether the conjectured problem is indeed hard. That amounts to either proving or disproving the conjecture, and both directions have resulted in very fruitful research.

**Some of the reduction mechanisms also turned out to be very powerful.**

Yes, that's another research direction. The idea that one can reduce this problem A to some other problem B sounds natural. However, it turns out that these reductions themselves are quite sophisticated, and to construct them, one needs a lot of new mathematical machinery. Once one develops that machinery, it becomes interesting on

its own, in the sense that now it doesn't even matter whether the conjecture itself is true or false.


**What are you working on these days?**

I have certainly been working on proving the conjecture, and in the last five years or so, with co-authors, I have proposed a possible plan toward proving it, so I'm excited about that. In particular, I've been working on the geometry of Grassmann graphs. These are a very specific class of graphs, and we want to understand their structural properties. In a recent paper with co-authors, we proposed how a better understanding of these graphs would make progress on the UGC.

**Are there any other things in the field going on that excite you, or directions you might move in the future?**

Broadly speaking, I'm very interested in the interaction between computer science and mathematics, especially geometry and analysis as mathematical areas, and there are certainly many things going on at this intersection. We currently have a large, collaborative project supported by the Simons Foundation on algorithms and geometry. Half of the researchers are from computer science, and half from math. So that's an ongoing project for the last three years that I'm very happy about.

**Can you talk about some of the work that's come out of the project?**

I can cite three striking results that show the back-and-forth interaction between computer science and geometry. I have been involved in work on the monotonicity testing problem in computer science and the related isoperimetric theorems on Boolean hypercube. Assaf Naor has been involved in work on the Sparsest Cut problem in computer science and the related questions about the geometry of Heisenberg group. Oded Regev has been involved in lattice-based cryptography and the related mathematical questions about integer lattices. In all these works, connections between CS and math have been discovered, benefiting both the fields, and it is difficult to say which inspired which. 

Leah Hoffmann is a technology writer based in Piermont, NY.

©2017 ACM 0001-0782/17/02

## Q&A

# Out of Bounds

*Mathematics led Subhash Khot, developer of the Unique Games Conjecture, to computer science without his ever having seen a computer.*

NEW YORK UNIVERSITY professor Subhash Khot has worked at the cutting edge of what cannot be done with computers since 2001 when, in his third year of graduate school at Princeton University, he formulated the groundbreaking Unique Games Conjecture (UGC). This seemingly simple statement—about the difficulty of solving a specific problem—turned out to have profound implications for the field. Khot has since received some of its highest honors, including the National Science Foundation's Alan T. Waterman Award, the International Mathematical Union's Rolf Nevanlinna Prize, and, most recently, the MacArthur Fellowship. Here, he recalls how it happened.

**Let's talk about your background. You grew up in India, and I understand you chose to study computer science without having seen a computer.**

It sounds quite strange, but that's how it was. I didn't have any exposure to computers or computer science, but I had very good exposure to mathematics in the form of specialized math exams and competitions throughout my school curriculum. And then I had to choose an undergraduate major, which in India one more or less has to declare before one starts the program. At that time, and probably this is the case even now, mathematics wasn't viewed as a good career option. But friends of mine told me that there are aspects of computer science that are really mathematical, and fortunately, that turned out to be the case.



**“In computer science, the best way to show that a certain problem is hard is to take another problem that is already known to be hard, and then reduce that problem to the first problem.”**

**After college, you went to Princeton for your Ph.D.**

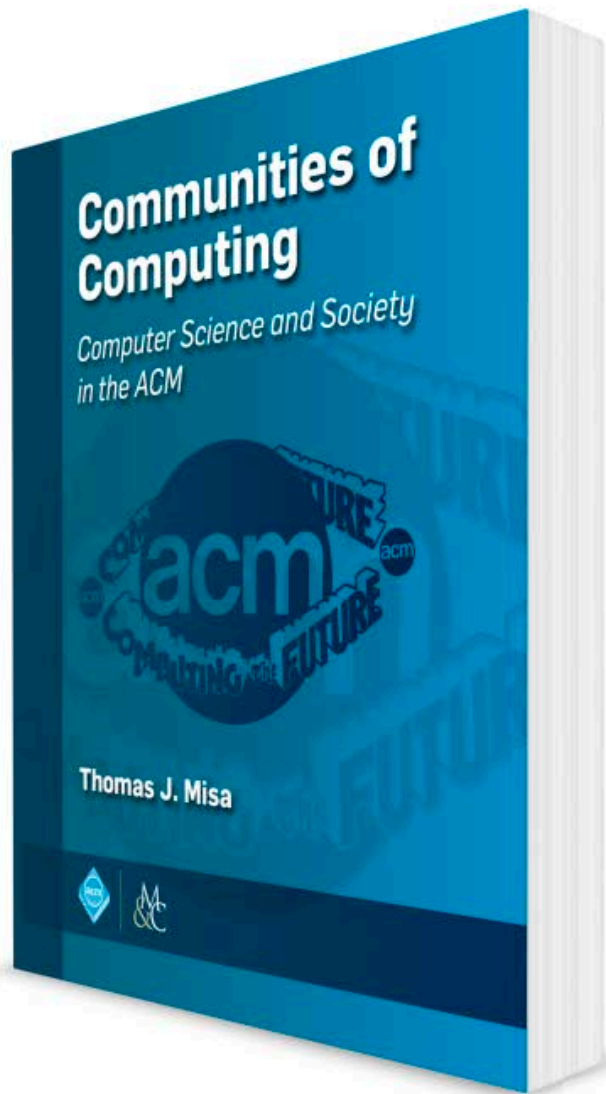
That decision was rather straightforward. I went to the Indian Institute of Technology in Bombay as an undergrad, and it was a very standard routine for most of the students to join Ph.D. programs in the U.S.

**Let's talk about the Unique Games Conjecture and how you came to develop it.**

In my early graduate school days, I was exploring different research topics. One of them was the hardness of approximation. There are many computational problems, known as NP-hard problems, that researchers believe are hard to solve. One can then ask whether one can solve them approximately. When an approximation is allowed, some problems do become easy. But others still remain hard, and the question is whether one can classify problems in terms of hardness of approximating them. This is the topic I started investigating, and in a couple of years, I was led naturally to the Unique Games Conjecture.

**It is also a topic that your advisor, Sanjeev Arora—who made extremely influential contributions toward proving the PCP Theorem—is well known for.**

Yes, my advisor Sanjeev Arora was involved in the pioneering work on this topic in the early 1990s, known as the PCP Theorem. By the time I started my graduate studies, which was in the early 2000s, large progress had been made, but the field was somehow stuck. [CONTINUED ON P. 103]



**Your first book-length  
history of the ACM.**

**Defining the Discipline  
Broadening the Profession  
Expanding Research Frontiers**

**Thomas J. Misa (Editor)**

*Charles Babbage Institute (University of Minnesota)*

The SIGs, active chapters, individual members, notable leaders, social and political issues, international issues, computing and community education...all are topics found within this first book-length history of the Association for Computing Machinery (ACM). Featuring insightful profiles of people who shaped ACM, such as Edmund Berkeley, George Forsythe, Jean Sammet, Peter Denning, and Kelly Gotlieb, and honest assessments of controversial episodes, this volume deals with compelling and complex issues involving ACM and computing.

This is not a narrow organizational history. While much information about the SIGs and committees are presented, this book is about how the ACM defined the discipline, broadened the profession, and how it has expanded research frontiers. It is a permanent contribution to documenting the history of ACM and understanding its central role in the history of computing.



ISBN: 978-1-970001-84-6 DOI: 10.1145/2973856

<http://books.acm.org>

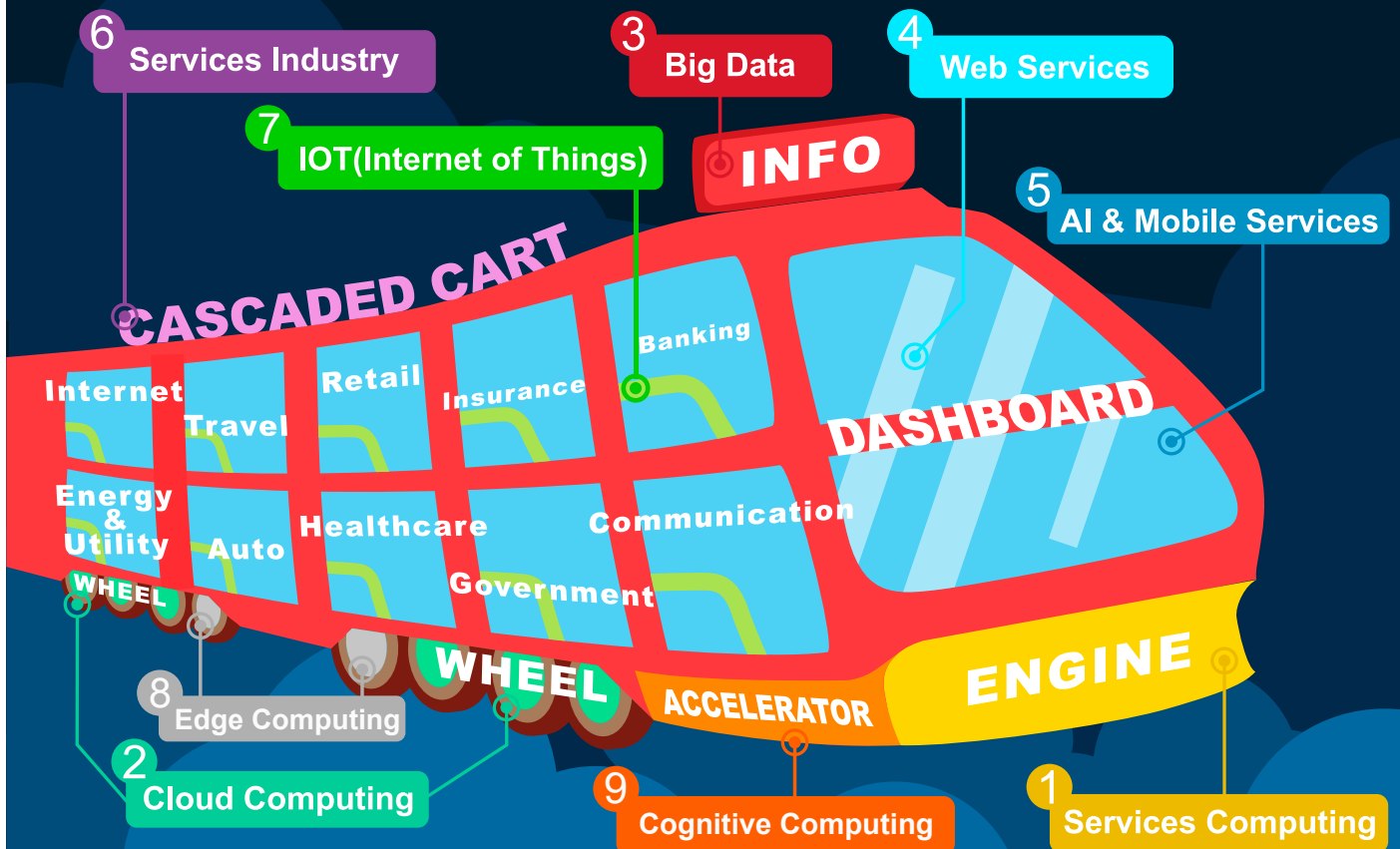
<http://www.morganclaypoolpublishers.com/misa>

# 2017 IEEE Services Congress IEEE BigData Congress

June 25 - June 30, 2017, Hawaii, USA



- ① IEEE 14th International Conference on Services Computing (SCC 2017)
- ② IEEE 10th International Conference on Cloud Computing (CLOUD 2017)
- ③ IEEE 6th International Congress on Big Data (BigData Congress 2017)
- ④ IEEE 24th International Conference on Web Services (ICWS 2017)
- ⑤ IEEE 6th International Conference on AI & Mobile Services (AIMS 2017)
- ⑥ IEEE 13th World Congress on Services (SERVICES 2017)
- ⑦ IEEE 2nd International Congress on Internet of Things (ICIOT 2017)
- ⑧ IEEE 1st International Conference on Edge Computing (EDGE 2017)
- ⑨ IEEE 1st International Conference on Cognitive Computing (ICCC 2017)



## Submission Deadlines

2/6/2017: ICWS 2017 (<http://icws.org>)  
 2/6/2017: CLOUD 2017 (<http://theCloudComputing.org>)  
 2/21/2017: SCC 2017 (<http://theSCC.org>)  
 2/21/2017: AIMS 2017 (<http://theMobileServices.org>)  
 2/28/2017: BigData Congress 2017 (<http://ieeeBigData.org>)  
 2/28/2017: SERVICES 2017 (<http://ServicesCongress.org>)

3/10/2017: ICIOT 2017 (<http://iciot.org>)  
 3/10/2017: EDGE 2017 (<http://theEdgeComputing.org>)  
 3/10/2017: ICC 2017 (<http://theCognitiveComputing.org>)

Contact: [confs@ServicesSociety.org](mailto:confs@ServicesSociety.org)

Dr. Liang-Jie Zhang (LJ), Steering Committee Chair

