

COMMUNICATIONS

CACM.ACM.ORG OF THE

ACM

01/2017 VOL.60 NO.01

Exponential Laws of Computing Growth

Bias in
Technology

Artificial
Intelligence:
Think Again

Cell-Graphs

Deploying SDN
in the Enterprise

Technology for
the Most Effective
Use of Mankind

MOORE'S
LAW

b.1965 -

REPORTS OF MY DEATH
ARE GREATLY EXAGGERATED

Association for
Computing Machinery



Systor2017

May 22 – 24 Haifa, Israel

10 The 10th ACM International
Systems and Storage Conference

Paper submission deadline: February 22, 2017

Paper acceptance notification: March 29, 2017

Poster submission deadline: March 22, 2017

We invite you to submit original and innovative papers, covering all aspects of computer systems technology, such as file and storage technology; operating systems; distributed, parallel, and cloud systems; security; virtualization; and fault tolerance, reliability, and availability. SYSTOR 2017 accepts full papers, short papers, and posters.

Program chairs

Peter Desnoyers, Northeastern University

Eyal de Lara, University of Toronto

General chair

Doron Chen, IBM Research

Posters chair

Adam Morrison, Tel Aviv University

Steering committee head

Michael Factor, IBM Research

Steering committee

Ethan Miller, University of California Santa Cruz

Liuba Shrira, Brandeis University

Dan Tsafir, Technion

Dalit Naor, IBM Research

Erez Zadok, Stony Brook University

Sponsored by



In cooperation with



Platinum sponsor



Gold sponsors



Sponsors



www.systor.org/2017/



Inviting Young Scientists



Association for
Computing Machinery

Meet Great Minds in Computer Science and Mathematics

As one of the founding organizations of the Heidelberg Laureate Forum <http://www.heidelberg-laureate-forum.org/>, ACM invites young computer science and mathematics researchers to meet some of the preeminent scientists in their field. These may be the very pioneering researchers who sparked your passion for research in computer science and/or mathematics.

These laureates include recipients of the ACM A.M. Turing Award, the Abel Prize, the Fields Medal, and the Nevanlinna Prize.

The Heidelberg Laureate Forum is **September 24–29, 2017** in Heidelberg, Germany.

This week-long event features presentations, workshops, panel discussions, and social events focusing on scientific inspiration and exchange among laureates and young scientists.

Who can participate?

New and recent Ph.Ds, doctoral candidates, other graduate students pursuing research, and undergraduate students with solid research experience and a commitment to computing research

How to apply:

Online: <https://application.heidelberg-laureate-forum.org/>
Materials to complete applications are listed on the site.

What is the schedule?

Application deadline—**February 14, 2017**.

We reserve the right to close the application website early depending on the volume

Successful applicants will be notified by **end of March/early April 2016**.

More information available on Heidelberg social media



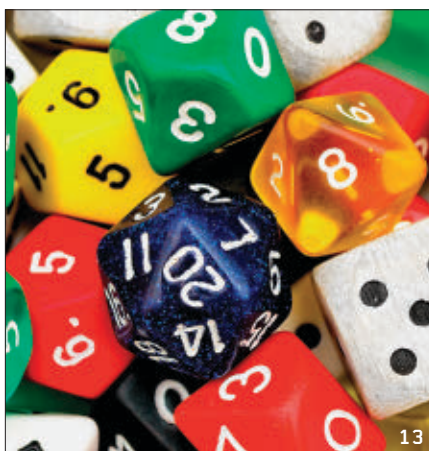
Departments

- 5 **Editor's Letter**
Technology for the Most Effective Use of Mankind
By Moshe Y. Vardi
-
- 7 **From the President**
The ACM Future of Computing Academy
By Vicki L. Hanson
-
- 9 **Cerf's Up**
Information and Misinformation on the Internet
By Vinton G. Cerf
-
- 10 **BLOG@CACM**
How We Teach CS2All, and What to Do About Database Decay
 Valerie Barr considers how to make computer science education meaningful and relevant to all, while a team from the Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory offers strategies to counter database decay.
-
- 25 **Calendar**
-
- 105 **Careers**

Last Byte

- 112 **Upstart Puzzles**
Open Field Tic-Tac-Toe
By Dennis Shasha

News



- 13 **Pure Randomness Extracted from Two Poor Sources**
 Developments from several disparate areas of computer science provide “a huge jump, both technically and also quantitatively.”
By Don Monroe
-
- 16 **Mapping the Internet of Things**
 Researchers are discovering surprising new risks across the fast-growing IoT.
By Alex Wright
-
- 19 **Bias in Technology**
 As leading companies release troubling diversity statistics, experts search for solutions.
By Gregory Mone

Viewpoints

- 22 **Technology Strategy and Management**
Is Google's Alphabet a Good Bet?
 A relatively simple query raises myriad complicated issues.
By Michael A. Cusumano
-
- 26 **Law and Technology**
Why Less Is More When It Comes to Internet Jurisdiction
 Considering legal uncertainty in the online environment.
By Michael Geist
-
- 29 **Historical Reflections**
Colossal Genius: Tutte, Flowers, and a Bad Imitation of Turing
 Reflections on pioneering code-breaking efforts.
By Thomas Haigh
-
- 36 **Viewpoint**
Artificial Intelligence: Think Again
 Social and cultural conventions are an often-neglected aspect of intelligent-machine development.
By Jerry Kaplan
-
- 39 **Viewpoint**
Effects of International Trafficking in Arms Regulations Changes
 Considering the impact of recent ITAR changes to the U.S. software industry and software education.
By Jeremy Straub



Association for Computing Machinery
 Advancing Computing as a Science & Profession

Practice



- 42 **Resolving Conflict**
Don't "win." Resolve.
By Kate Matsudaira

- 45 **Faucet: Deploying SDN in the Enterprise**
Using OpenFlow and DevOps for rapid development.
By Josh Bailey and Stephen Stuart

- 50 **Research for Practice: Web Security and Mobile Web Computing**
Expert-curated guides to the best of CS research.



Articles' development led by **acmqueue**
queue.acm.org

Contributed Articles



- 54 **Exponential Laws of Computing Growth**
Moore's Law is one small component in an exponentially growing planetary computing ecosystem.
By Peter J. Denning and Ted G. Lewis



Watch the authors discuss their work in this exclusive *Communications* video.
<http://cacm.acm.org/videos/exponential-laws-of-computing-growth>

- 66 **Bottom-Up Enterprise Information Systems: Rethinking the Roles of Central IT Departments**
Central IT needs to guide functional areas and departments toward effective operational and procurement practices.
By Cecil Eng Huang Chua and Veda C. Storey

Review Articles

- 74 **Cell-Graphs: Image-Driven Modeling of Structure-Function Relationship**
Cell-graph construction methods are best served when physics-driven and data-driven paradigms are joined.

By Bülent Yener



Watch the author discuss his work in this exclusive *Communications* video.
<http://cacm.acm.org/videos/cell-graphs>

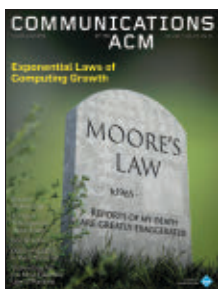
Research Highlights

- 86 **Technical Perspective**
Magnifying Motions the Right Way
By Richard Szeliski

- 87 **Eulerian Video Magnification and Analysis**
By Neal Wadhwa, Hao-Yu Wu, Abe Davis, Michael Rubinstein, Eugene Shih, Gautham J. Mysore, Justin G. Chen, Oral Buyukozturk, John V. Guttag, William T. Freeman, and Frédo Durand

- 96 **Technical Perspective**
Mapping the Universe
By Valentina Salapura

- 97 **HACC: Extreme Scaling and Performance Across Diverse Architectures**
By Salman Habib, Vitali Morozov, Nicholas Frontiere, Hal Finkel, Adrian Pope, Katrin Heitmann, Kalyan Kumaran, Venkatram Vishwanath, Tom Peterka, Joe Insley, David Daniel, Patricia Fasel, and Zarija Lukić



About the Cover:

Peter Denning and Ted Lewis take a new look at Moore's Law and the causes of exponential growth for information technologies and find there is much more life left in the landmark observation. Cover illustration by Peter Crowther Associates.



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 Guerrin; 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

1828 L Street, N.W., Suite 800

Washington, DC 20036 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

Juliet Chance

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

acmhlp@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

Juliet Chance

acmm mediasales@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

eric@cacm.acm.org

NEWS

Co-Chairs

William Pulleyblank and Marc Snir

Board Members

Mei Kobayashi; Michael Mitzenmacher;

Rajeev Rastogi

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

Q 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 *Queue*.

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; Amir El Abbadi; Sanjeev Arora;

Michael Backes; Nina Balcan; Dan Boneh;

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.; David Wagner;

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 acmhlp@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





DOI:10.1145/3020075

Moshe Y. Vardi

Technology for the Most Effective Use of Mankind

TECHNO-OPTIMISM IS DEFINED AS the belief that technology can improve the lives of people. It was famously satired in the U.S. television comedy series “Silicon Valley,” with a startup-company’s founders pledging to “make the world a better place through Paxos algorithms for consensus protocols.” But some people take techno-optimism very seriously. Ray Kurzweil, an accomplished tech innovator, described his techno-optimistic vision in his books: *The Age of Spiritual Machines*, *How to Create a Mind: The Secret of Human Thought Revealed*, and *The Singularity Is Near*.

In a keynote address (see <https://goo.gl/RwkwK1>) at the 2016 meeting of the Computing Research Association, Ken- taro Toyama argued that “In spite of the do-gooder rhetoric of Silicon Valley, it is no secret that computing technology in and of itself cannot solve systemic social problems.” Toyama’s argument, detailed at length in his 2015 book *Geek Heresy: Rescuing Social Change from the Cult of Technology*, is that persistent societal challenges, such as socio-economic inequality, do not have technology-centric solutions. Indeed, over the past 50 years we have witnessed the development of the Internet, the personal computer, the cellphone, the Web, search engines, social media, and smartphones—a development often summarized as the “Information Revolution.” During this period, the U.S. poverty rate oscillated in the 13%–15% range, completely impervious to developments in information technology. In view of the data on poverty, the quote attributed to Mark Zuckerberg, Facebook’s founder and CEO—“The richest 500 million [people] have way more money than the next six billion combined. You solve that by getting every-

one online”—is either hopelessly naïve or utterly self-serving.

Over the past decade alone we have witnessed the demise of two prominent techno-optimist “solutions.” The One Laptop per Child (OLPC) project was launched in 2005 with the goal of transforming education for the world’s disadvantaged schoolchildren. But within a few years “the vision was overwhelmed by the reality of business, politics, logistics, and competing interests worldwide” (see <https://goo.gl/xWj8OK>). The MOOC (massive open online courses) movement was launched in 2011 with the rhetoric of dramatically reducing the cost of higher education while “reaching the quality of individual tutoring.” But by 2014, Sebastian Thrun, a MOOC pioneer, concluded that “The basic MOOC is a great thing for the top 5% of the student body, but not a great thing for the bottom 95%.”

The central thesis of Toyama’s talk was that “Technology has positive impact *only when* amplifying social trends or institutions that are already positively inclined.” As much as I sympathize with Toyama’s skeptical approach toward techno-optimism, I find this thesis hard to swallow. Consider, for example, Polio-myelitis, often called polio, as an example. Polio used to be a dreaded infectious childhood disease, as in a small fraction of cases the disease results in permanent severe muscle weakness. In 1952, Jonas Salk developed the first effective polio vaccine, which led to drastic reduction in polio infections (100 known cases worldwide in 2015). Surely this counts as technology with positive impact. Of course one could argue the worldwide adoption of polio vaccination required “social trends or institutions that are already positively inclined,” but that would make the thesis a tautology. Some social problems do have technical solutions!

But Toyama is right that using technology to solve societal challenges requires a deep understanding of the societal context, an understanding that was not demonstrated by the OLPC and MOOC movements. Furthermore, deploying technology without understanding its societal context may have adverse societal consequences. Consider “frictionless sharing” as an example. The concept first arose in the context of scholarly work, where the goal was to enable scholars to easily share resources with other scholars. In 2011, Zuckerberg announced developments to Facebook that would allow “real-time serendipity in a frictionless experience.” By 2016, however, frictionless sharing has given rise to the *fake-news* phenomenon, with the proliferation of websites that publish fraudulent misinformation, quickly spread via social media, intended to mislead readers. While it is difficult to gauge the impact of this phenomenon on the outcome of the 2016 U.S. presidential election, the explosion of misinformation surely counts as a negative consequence of technology. Frictionless sharing is technology, but to what end?

We, as computing professionals, are engaged in the development of information technology. This technology is changing the world, but not always for the better. It is time for computing to emerge from its technological cocoon and engage vigorously with social science. If we wish our technology to be developed “for the most effective use of mankind” (quoting from Ada Lovelace’s 1843 letter to Charles Babbage), then we need to understand mankind better!

Follow me on Facebook, Google+, and Twitter.

Moshe Y. Vardi, EDITOR-IN-CHIEF

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>



DOI:10.1145/3020077

Vicki L. Hanson

The ACM Future of Computing Academy

One of my priorities as ACM president is to have our organization effectively engage with younger practitioners, researchers, educators, and entrepreneurs across

the global computing community. In my first column last August, I noted that new supporting initiatives were being formulated. I am now pleased to announce the establishment of the ACM Future of Computing Academy (ACM-FCA), created to bring together talented young professionals from various computing disciplines to address the most pressing challenges facing the field and society at large.

ACM's longtime members and our traditional activities have helped grow our field and have contributed to the enormous impact computing has made on our lives. We will look to the ACM-FCA members to lead the way in showing us how we might develop new models for participation, collaboration, and career support. In short, members of the Academy will engage in activity for the benefit of their own and future generations.

When I took office in July, I asked Matthias Kaiserswerth (Hasler Stiftung, Switzerland) and Aaron Quigley (University of St. Andrews, Scotland) to consider how ACM might better engage young computing professionals. They have taken on this challenge and conducted a wide-ranging consultation with existing academies, ACM executives, ACM Council, Europe Council, ACM SIGs, and others. Based on their recommendations, the ACM Council has voted to establish this new Future of Computing Academy.

The FCA will seek to harness collective action to define and launch new

ACM 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 all regions, computing disciplines, and computing professions. Academy members will be supported by an extended network of more senior mentors, ACM leadership, and recognized thought leaders in computing.

Given the FCA goals, we anticipate that members of the Academy will reflect the global diversity in computing and typically be in their 20s to early 30s upon application to the Academy. Members will have already demonstrated great professional promise, and should be eager to come together to engage in this work. Members of the academy are expected to participate in activities for the

The FCA will seek to harness collective action to define and launch new ACM initiatives that will carry us into the future.

benefit of their generation as they build and shape new experiences to advance computing.

If you are a young computing professional, you may recognize the opportunity that joining this Academy presents. Perhaps your current work as a researcher, practitioner, entrepreneur, or educator has given you insights into what problems ACM should be tackling, how the field of computing should evolve, how ACM should adapt to the needs of future generations, and where computing can help address the global issues we all face. If so, and if you have the desire to come together with like-minded individuals from around the world, then this Academy is for you.

Or perhaps you are a more established computing professional. If you wish to support this goal of developing fresh perspectives, then consider encouraging suitable candidates to apply.

The ACM-FCA will be taking applications at <http://www.acm.org/fca> through 15 March of 2017, with Matthias Kaiserswerth and Aaron Quigley serving as the initial convening committee. The Academy will be self-governing and will establish an elected executive committee.

It is important to note that membership in the Academy is a service commitment, not an award. Active participation is essential to Academy membership. The inaugural meeting of the ACM-FCA will be on 25 June 2017 in San Francisco. Members of the Academy will also be invited to attend the "50 Years of the ACM Turing Award" conference that precedes this meeting.

We look forward to the establishment of the Academy and to welcoming its first cohort of members in 2017. 

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.

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

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





DOI:10.1145/3018809

Vinton G. Cerf

Information and Misinformation on the Internet

In June 2016, the U.K. held a referendum on its membership in the European Union. In November 2016, the U.S. held its national elections. In the run-up to both of

these important decisional events, the Internet with its burgeoning collection of “information” dissemination applications, influenced the decisions of voters. The disturbing aspect of these (and many other decisional events) is the quantity of poor-quality content, the production of deliberately false information, and the reinforcement of bad information through the social media.

One reaction to bad information is to remove it. That’s sometimes called censorship although it may also be considered a responsible act in accordance with appropriate use policies of the entities that support information dissemination and exchange. A different reaction is to provide more information to allow viewers/readers to decide for themselves what to accept or reject. Another reaction is to provide countervailing information (fact checking) to help inform the public. Yet another reaction is simply to ignore anything that you reject as counter to your worldview. That may lead to so-called *echo chamber* effects where the only information you really absorb is that which is consistent with your views, facts notwithstanding.

The wealth (I use this word gingerly) of information found on the Internet is seemingly limitless. On the other hand, it is of such uneven quality that some of us feel compelled to exercise due diligence before accepting anything in particular. That calls for critical thinking and, as I have written in the past, this is something that not everyone is prepared to or willing to expend energy on. That is

not a good sign. A society that operates on the basis of bad or biased information may soon find itself in difficulties because decisions are being made on shaky ground.

Unfortunately, we don’t seem to be able to guarantee that decision makers, including voters, will apply critical thinking, due diligence, and fact checking before taking decisions or propagating and reinforcing bad quality or deliberately counterfactual information. While the problem is more recognized now than ever, the proper response is far from agreed upon. It may even prove necessary to experiment with various alternatives. For example, rumors propagate rapidly through social media and recipients need tools to debunk them. The SNOPE website (www.snopes.com) provides information to expose false rumors or to confirm them using factual information and analysis. We can use more of this.

Of course, in many cases, the situation is less clear-cut and differences of opinion illustrate that there can be conflicting views of truth or falseness. What seems important is to have access to as much factual information as possible and to distinguish that from the opinions about the implications of these facts. U.S. politician Daniel Moynihan is credited with the observation that you are not entitled to your own facts, only to your opinions. Even here, of course, one can encounter differences of opinion about what is factual and what is not.

This suggests to me that in the modern Internet environment, where any-

one can say pretty much anything and others can read it, we are in need of processes that will help readers/viewers who wish to evaluate for factual value what they see and hear. It is notable that in the waning period of the political campaigns leading up to the U.S. presidential election, some media began providing fact-checking to go along with their reporting. The malleability of content on the Internet and its potentially ephemeral nature reinforces my belief that history is important and that its preservation is an important part of democratic societies.

This leads me to conclude that ways to preserve the content of the Internet in the interest of avoiding revisionist history may prove to be an important goal for technologists who worry about these things. This must be balanced against notions such as “the right to be forgotten” that are emerging in various jurisdictions, most notably in the European Union. There are legitimate reasons to remove harmful information that makes its way onto the Internet, such as child pornography and information that leads to identity theft, for example. Finding a balance that preserves the value of historical record, corrects false or incorrect information, and supports due diligence and critical thinking is a challenge for our modern information era. 

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

Copyright held by author.

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:10.1145/3014349

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

How We Teach CS2All, and What to Do About Database Decay

Valerie Barr considers how to make computer science education meaningful and relevant to all, while a team from the Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory offers strategies to counter database decay.



Valerie Barr
How We Teach Should Be Independent of Who We Are Teaching
<http://bit.ly/2eYnx0Z>
October 11, 2016

For many years I have been part of discussions about how to diversify computing, particularly about how we recruit and retain a more diverse cohort of computer science (CS) students. I wholeheartedly support this goal, and spend a considerable amount of my effort as chair of ACM-W helping to drive programs that focus on one aspect of this diversification, namely encouraging women students to stay in computing.

Of late I have become very concerned about how some elements of the diversity argument are being expressed and then implemented in teaching practices. A shorthand has developed that often comes out as two problematic claims:

Problem 1. Women are motivated by social relevance, so when we teach

them we have to discuss ways in which computing can contribute to the social good.

Problem 2. Students from underrepresented minorities (URM) respond to culturally relevant examples, so when we teach them we have to incorporate these examples into course content.

This formulation of what we should be doing in the classroom is problematic for a number of reasons:

1. These statements are silent on the subject of white and Asian men, the groups that dominate in CS classrooms, effectively implying that these people are not interested in computing for the social good or culturally relevant examples, that they are only motivated by the hard-core geeky techie parts of computing.

2. This formulation paints all women with a single brush, and does the same for URM students. Some women are interested in the social relevance of computing, but are all women going to be motivated by this? Some URM

students are motivated by culturally relevant examples, but are all URM students going to be motivated by this?

3. While painting women and URM with a single brush, this formulation implies that members of these groups are not interested in computing for techie reasons, that members of these groups will not ever be excited about the technology in its own right.

4. Further, there is an implication that we need to discuss the social relevance of computing *only* when there are women in the class, and we need to utilize culturally relevant examples *only* when there are URM students in the class.

5. The logical, and dangerous, final conclusion is that if there are *only* Asian and white men in the room, then we do not need to make any changes at all to course content or pedagogy.

These assumptions about students can have a very negative impact on our teaching, causing us to potentially drive away the very students we are hoping to recruit and retain. As we continue efforts to diversify computing, we cannot afford to paint *any* group in a monochromatic way. We have to embrace the richness of today's student population by making what we teach meaningful and relevant to them. There are women who want to geek out about hard-core tech, and there are men who care deeply about computing for the social good. There are students of all genders and ethnic and racial backgrounds who will be happy with an old-fashioned lecture, and those who will thrive on

active learning with examples drawn from a range of cultures and application areas. Many students will be motivated by knowing how the techniques and subject matter they are learning fit into their future workplace or life goals.

In order to change the toxic climate in tech, a climate that, for example, leads 45% of women to leave tech jobs within five years, we have to teach *everybody* differently. If we pretend that all women students are the same, and all URM students are the same, and all Asian and white male students are the same, then we will never adequately address the blind spots and weaknesses in instruction and curriculum development that have led to our current situation. A rich approach to curriculum and teaching pedagogy will maximize our ability to reach *all* kinds of learners, all parts of the student population. We have to use varied content and pedagogies regardless of whom we see in the room and work to connect to what students know or care about. This approach will guarantee that all students, including those from the groups that currently dominate computing, will be exposed to a rich, multifaceted, view of computing, be better equipped to address the challenges of the field, and be better equipped to work collegially within a diverse workforce.

Thanks to several colleagues who gave me important feedback on prior versions of this post.



**Michael Stonebraker,
Raul Castro Fernandez,
Dong Deng, and
Michael Brodie**
**Database Decay and
What To Do About It**

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

October 24, 2016

The traditional wisdom for designing database schemas is to use a design tool (typically based on a UML (https://en.wikipedia.org/wiki/Unified_Modeling_Language) or ER (https://en.wikipedia.org/wiki/Entity-relationship_model) model) to construct an initial data model for one's data. When one is satisfied with the result, the tool will automatically construct a collection of 3rd normal form relations for the model. Then, applications are coded against this rela-

tional schema. When business circumstances change (as they do frequently), one should run the tool again to produce a new data model and a new resulting collection of tables. The new schema is populated from the old schema, and the applications are altered to work on the new schema, using relational views whenever possible to ease the migration. In this way, the database remains in 3rd normal form, which represents a "good" schema, as defined by DBMS researchers. "In the wild," schemas often change once a quarter or more often, and the traditional wisdom is to repeat the above exercise for each alteration.

In a survey of 20 database administrators (DBAs) at three large companies in the Boston area, we found that this traditional wisdom is rarely-to-never followed for large, multidepartment applications. Instead, DBAs try very hard not to change the schema when business conditions change, preferring to "make things work" without schema changes. If they must change the schema, they work directly from the relational tables in place. Using these tactics, the ER or UML model (if it ever existed) diverges quickly from reality. Moreover, over time, the actual semantics of the data tend to drift farther and farther from a 3rd normal form data model.

We term this divergence of reality from 3rd normal form principles **database decay**. Over time, decay becomes worse and worse, leading to **rotted** databases and ultimately to databases that are so decayed that they cannot be further modified. Obviously, this is a very undesirable state of affairs.

In our opinion, the reason for decay stems from the multidepartment organization of large implementations. Hence, various pieces of the overall application are coded by different organizations, typically using ODBC (https://en.wikipedia.org/wiki/Open_Database_Connectivity) or JDBC (https://en.wikipedia.org/wiki/Java_Database_Connectivity) to specify the SQL in transactions. If one business unit needs to change the semantics of the database, it is exceedingly difficult to figure out what code from other departments needs to be changed and how extensive the required repairs are. In our opinion, this leads DBAs to change the schema in such a way that application maintenance is minimized,

rather than making a change that maximizes the cleanliness of the data. Of course, the result of a different DBA cost function is database decay and rot.

Seemingly, database decay is a fact of life, which ultimately renders databases unable to be modified. There are three strategies that can counter database decay.

The first one is to construct **defensive schemas** in the first place. Such schemas are more resilient to subsequent changes than ones produced using the traditional wisdom. We have developed a methodology for such schemas, which will be addressed in an upcoming paper.

The second tactic is to write **defensive application code**. For example, one should never use `Select * From Table-name`, because it tends to make applications break if attributes are added or deleted downstream.

Lastly, in our opinion, it is a bad practice to let application groups directly code against an ODBC/JDBC interface. This is what is responsible for DBAs not knowing the impact of possible schema changes. Instead, we advocate requiring application groups to use a messaging interface to send higher-level commands to a DBMS. These messages are intercepted and turned into SQL in server-side code. Such an architecture localizes DBMS code that may need to be changed later on. Moreover, we have written a prototype system that can examine such code and determine if it needs to be changed as a result of schema evolution. In this way, we expect to lower the cost of schema changes, and perhaps slow down or arrest database decay. An upcoming paper details our prototype.

We are looking for "in the wild" database projects that are dealing with schema evolution that would be amenable to trying our prototype system. If you are interested, please contact Michael Brodie at mlbrodie@mit.edu. 

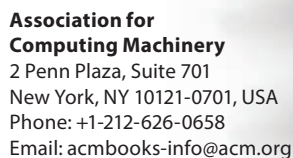
Valerie Barr is a professor in the Computer Science department of Union College, and serves as chair of ACM-W, the ACM Council on Women in Computing. **Michael Stonebraker** is an adjunct professor in the Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory (CSAIL), and recipient of the 2014 ACM A.M. Turing Award. **Michael Brodie** is a Research Affiliate in CSAIL, while **Dong Deng** and **Raul Castro Fernandez** are post-doctoral researchers at CSAIL.

© 2017 ACM 0001-0782/17/1 \$15.00

ACM Books.

For more info please visit
<http://books.acm.org>

For more info please visit



**Association for
Computing Machinery**

Phone: +1-212-626-0658

Email: acmbooks-info@acm.org



**Morgan & Claypool
Publishers**

San Rafael, CA 94901, USA

Phone: +1-415-462-0004

Email: info@morganclaypool.com



Pure Randomness Extracted from Two Poor Sources

Developments from several disparate areas of computer science provide “a huge jump, both technically and also quantitatively.”

TRULY RANDOM NUMBERS are critical for computing and cryptography. Although deterministic algorithms can generate numbers that seem random, critical applications depend upon truly unpredictable physical processes. Unfortunately, the resulting sequences can still contain hidden regularities, so theoretical computer scientists have long sought “extractors” that produce perfect randomness from these imperfect sources.

It has been known for 30 years that there should be many ways to combine two imperfect sources to generate nearly perfectly random numbers, but only now have researchers shown explicitly how to create such “two-source extractors.” The work, which received a best-paper award at the 48th ACM Symposium on Theory of Computing (STOC 2016) last June, brought together developments from several disparate areas of computer science. “The search has been going on mainly in the past 10 or 15 years, and this is the culmination of this effort,” said Avi Wigderson of the Institute for Advanced Study at



Princeton University in Princeton, NJ. “It’s a huge jump, both technically and also quantitatively.”

The practical implications may be modest for now, though. Security specialist Bruce Schneier, for one, does not see any urgent need for better random numbers. Schneier helped create the widely used Fortuna pseudorandom-number generator, and says there are

already many adequate sources. “In my world no one’s worried about this,” he said. “These systems already work” to provide secure communications when attention is paid to all of the other implementation details. “We have lots of problems; this isn’t one of them.”

Wigderson’s enthusiasm is more fundamental. “The point of studying these things theoretically is you’d like

guarantees about the behavior of a published algorithm or cryptographic protocol.” Moreover, other researchers have already built on the results to devise even more efficient algorithms. As a side benefit, the two-source extractor also has implications for other mathematical issues, including solving a 70-year-old problem in graph theory.

An Adversarial Approach

Random numbers are used in a wide variety of computational tasks. For example, the outcome of complex processes that cannot be calculated explicitly, such as climate change, can be modeled by averaging random samples from a probability distribution. Other problems are just easier to attack using random numbers. In fact, some problems have proven-efficient solutions based on random numbers, while the best deterministic algorithms known require impractically long calculations.

Theoretically, assessment of these methods is only possible if the numbers are really random, meaning that every outcome is equally likely, no matter what has come before. Any hidden regularities could introduce biases that skew the results, and there are published cases that suffer from such distortions, said David Zuckerman of the University of Texas, Austin, who coauthored the new work with graduate student Eshan Chattopadhyay, now at the Institute for Advanced Study.

To guard against such problems, “in the computer science setting you try to think as adversarially as possible,” said Henry Yuen of the University of California, Berkeley. “You try to think that nature is conspiring against you.” This adversarial approach is particularly appropriate in cryptography, where randomly generated keys, shared between sender and legitimate recipient, are used to hide information from potential eavesdroppers. “What you’d really like to be sure is that there’s no other being in the universe that can predict the outcome,” Yuen said. “It might be uncertain to you, but how do you know it’s uncertain to someone else?”

For example, if strings of 100 bits were chosen randomly, any particular string would only occur about every 2^{100} (about 1,000,000,000,000,000,000,000,000,000) times. If a particular string occurred much more often, say

It turns out to be surprisingly difficult to extract guaranteed high-quality randomness from a single source, if the source is even modestly non-random.

every 1,000 times, a human observer would certainly not notice, unless the string itself was distinctive. Moreover, a statistical test could miss this rare violation of randomness unless it was told to test for the specific string. A malevolent adversary who knew about that special string, however, could use that knowledge to drastically reduce the computational challenge of deducing an encoded signal.

Quantifying Randomness

Assessing deviations from randomness can be subtle. Theoretical computer scientists use a quantity called min-entropy, which is the negative base-2 logarithm of the highest probability of occurrence out of all possible sequences. (The more-familiar Shannon entropy quantifies the information in a message as the *average* of this logarithm over all sequences, which is always larger than this worst case.) In the one-in-a-thousand example discussed earlier, the min-entropy would be $\log_2(1,000)$, or about 10 bits, which is one-tenth of the 100 bits of a truly random distribution.

One way to get sequences with near-maximal min-entropy is to use pseudorandom number generators, which use deterministic algorithms to create sequences that are mathematically assured of appearing random. For example, some algorithms cycle irregularly through all possible strings starting from some “seed” sequence. For a particular seed, however, this pseudorandom sequence will be the same. This repeatability can be convenient for debugging code, but creates vul-

nerabilities, so cryptography protocols like Fortuna spend a lot of effort on frequently injecting new seeds.

To completely avoid pseudorandomness, computer chips often come with physical random number generators based on the outcome of noisy electronic events. There are even companies that sell devices based on the outcome of quantum processes that are not predictable, even in principle.

These physically derived sources may still contain hidden regularities, however, for example because of correlations between different events. The quality of the randomness also depends on how manufacturers have implemented all of the other details. Some researchers have even speculated, especially since the disclosures by Edward Snowden, that commercial hardware sources could be subtly manipulated to make the codes easier to break by the U.S. National Security Agency or others. Researchers, therefore, would like to have assured randomness even from these sources.

Harvesting Randomness

It turns out to be surprisingly difficult to extract guaranteed high-quality randomness from a single source, if the source is even modestly non-random. For some 30 years, theoretical computer scientists have recognized and pursued two related ways to circumvent this problem. Both cases had been proved to have appropriate algorithms; the problem has been to find them.

The first approach leverages a small, high-quality random seed to extract high-quality random bits from a poor source. For a decade now, researchers have known how to efficiently construct such “seeded extractors” that are essentially as efficient in making use of low-quality sources as is theoretically possible; if the source is an n -bit string, for example, that string need only have a min-entropy of order $\log(n)$.

The second approach avoids the need for a perfect seed by combining two low-quality sources, which must be independent. Researchers had long ago proved that almost all possible extraction algorithms would work. “It’s like finding hay in a haystack,” Zuckerman joked. Yet in fact, for a decade now, the best-known two-source extractor still required that each n -bit source string have a min-entropy of almost $n/2$ bits.

“A lot of problems in theoretical computer science are like this,” Zuckerman said, where the challenge is finding explicit examples of entities that are known to exist. Using a complex procedure, Chattopadhyay and Zuckerman showed how to construct extractors that require only $\log(n)$ to a power, specifically the 74^{th} power.

An early draft of the paper was posted online in the summer of 2015. “This was such a big result that all the experts immediately just started reading it. It quickly became clear that this was a breakthrough result and all the ideas were correct—and also quite clever,” said Yuen. Other researchers quickly extended the results, reducing the exponent from 74 to slightly over one. Zuckerman’s former student Xin Li, now at Johns Hopkins University, also adapted the method to generate multiple bits, as opposed to the single bit of the original paper.

“Everyone has been waiting for this,” said Li, but he cautioned that “most of this work is still theoretical.” Li said he expects the work will eventually find its way into practical computing use, but “right now the running time is not very good, so it will take some time to improve this.”

Surprising Connections

To explicitly construct two-source extractors, Chattopadhyay and Zuckerman brought together three sets of “tools that were lying around the garage,” but which had been developed for seemingly unrelated problems, Wigderson said. “The combination is really pretty amazing.”

The first tools, called “non-malleable extractors,” are similar to seeded extractors, and had been studied for cryptography. Xi Lin had recently highlighted their unexpected usefulness in the two-source problem. The second tool, drawing on decade-old work on pseudorandom generators for parallel computing, also had “seemed totally irrelevant” to the current problem, Wigderson said. Finally, Chattopadhyay and Zuckerman extended 30-year-old work to explicitly construct resilient functions, which can ensure that voting is not swayed by small coalitions of voters (or adversarially selected bits).

“Part of our contribution was getting a way to convert two low-quality sources and combine them in a certain

way so that most of the bits are random but there are a few that are not,” Zuckerman said. The resilient function then ensures that the few non-random bits cannot prejudice the outcome.

An important by-product of the two-source extractor design, Wigderson said, is an explicit solution of another longstanding challenge in graph theory. Specifically, it concerns Ramsey graphs, graphs that contain no large cliques of connected nodes or groups of unconnected nodes larger than a certain size.

“It’s amazing, but if you would like to build a large graph with this property, it’s very hard,” Wigderson said, even though almost all randomly chosen graphs have it, as shown statistically by the prolific mathematician Paul Erdős in 1947. “People have struggled ... for many, many decades” to construct them systematically, he said.

The tools developed by Chattopadhyay and Zuckerman show how to create such graphs systematically. (Independently, Gil Cohen, now at Princeton University, also presented a solution to this problem at STOC 2016.) “The general quest” of dealing with randomness in combinatorial situations, Wigderson said, “is a big deal in many ways.” **C**

Further Reading

Chattopadhyay, E., and Zuckerman, D., Explicit Two-Source Extractors and Resilient Functions, *Electronic Colloquium on Computational Complexity*, Revision 2 of Report No. 119 (2015), <http://eccc.hpi-web.de/report/2015/119/>

Chattopadhyay, E., and Zuckerman, D., How random is your randomness, and why does it matter? *TheConversation.com*, Sept. 18, 2016, <https://theconversation.com/how-random-is-your-randomness-and-why-does-it-matter-59958>

Recommendation for the Entropy Sources Used for Random Bit Generation, (Second draft January 2016), National Institute of Standards and Technology Special Publication 800-90B, <http://csrc.nist.gov/publications/PubsDrafts.html#800-90B>

Wigderson, A., Center for the Study of Rationality, Randomness. <https://www.youtube.com/watch?v=syUxHJFwwQQ>

Don Monroe is a science and technology writer based in Boston, MA.

© 2017 ACM 0001-0782/17/1 \$15.00

ACM Member News

A WINDING ROAD TO SENTIMENT ANALYSIS



Bing Liu, a professor of computer science at the University of Illinois at Chicago, is a

recognized expert in determining whether online reviews—a prime driver in the e-commerce world—are genuine. However, he did not follow a straight path in developing his expertise in opinion mining.

Liu earned an undergraduate degree in mechanical engineering from the Northwest Institute of Light Industry, Shaanxi, China. Growing more intrigued by artificial intelligence (AI), Bing went to the University of Edinburgh, U.K., earning his Ph.D. in that discipline in 1989. He spent several years at the Information Technology Institute, the applied research and development arm of the National Computer Board (now the Infocomm Development Authority of Singapore), before becoming a research fellow at the National University of Singapore, where he became a professor in 1995.

“I came to the U.S. in 2002, and have been at the University of Illinois in Chicago since then.”

“Around 1996 I got into data mining, and mining text on the Web in 2001,” says Bing, “and that is the beginning of my interest in sentiment analysis,” extracting information on opinions, emotions, and sentiments from text.

Bing cites online reviews as substantive sources of information on social media and the Internet. “When someone writes one, are they positive or negative? What is the sentiment? What emotions are being expressed? I got interested in analyzing these things from text.”

He is now particularly interested in machine learning research as applied to sentiment analysis and opinion mining, and sees chatbots as the next frontier in analyzing online sentiments.

—John Delaney

Mapping the Internet of Things

Researchers are discovering surprising new risks across the fast-growing IoT.

AS MORE AND more physical objects get connected to the Internet—from consumer products like webcams and pacemakers to industrial equipment like wind turbines and power plants—the contours of the Internet are shifting beyond the realm of screen-based devices to encompass a much broader swath of the world around us.

Wherever the Internet goes, security risks seem to follow. As the Internet of Things (IoT) continues to expand, those risks are taking on new dimensions well beyond the familiar threats of stolen passwords and credit cards.

“When you say ‘Internet of Things,’ the first thing most people think of are Apple Watches or Fitbits,” says David O’Brien, a senior researcher at Harvard University’s Berkman-Klein Center for the Internet and Society.

“They’re not thinking about programmable logic controllers or other infrastructure devices.”

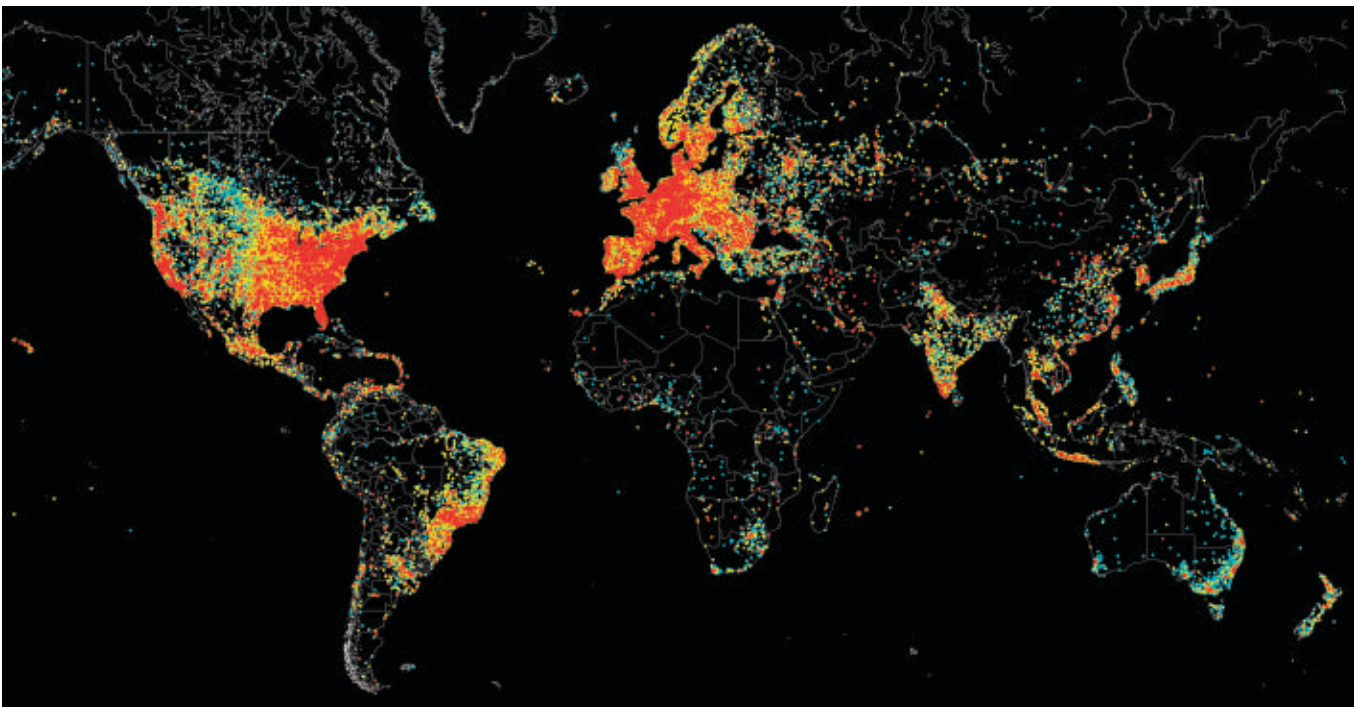
Industrial computing devices are a vast, largely invisible realm of the IoT, one that remains out of sight to most of us, yet plays a critical role in sustaining our everyday quality of life: power plants, water pumps, and oil rigs all rely on industrial computers connected to the Internet, and these devices appear to be far less secure than we might assume.

The lack of security across the industrial IoT has come to light largely thanks to an experimental search engine called Shodan. First launched in 2009, the service now crawls nearly four billion devices over the IPv4 network, as well as a number of IPv6-connected devices. At any given time, it monitors about 700 million devices (depending on network connectivity, and whether the devices are turned off or on).

Shodan’s creator, John Matherly, first started work on the service as a teenager in the mid-2000s. “The idea for Shodan came to me during the age of peer-to-peer software such as Napster and E-mule. The original concept was to provide a tool that would let security researchers scan networks and share the data (via P2P) with others.”

Unlike Web browsers that traverse the Internet via the Hypertext Transport Protocol (HTTP), Shodan surveys other TCP/IP-connected ports including FTP, SSH, SNMP, SIP and RTSP ports in search of responsive servers. When it receives a welcome message (or any response), it retrieves what metadata it can find, and catalogs the information.

At first, Matherly envisioned collecting data on the kinds of Internet-connected products in use, and create a repository of information about patches, site licenses, and other useful meta-



Shodan founder John Matherly used the search engine to map all Internet-connected device in the world.

IMAGE COURTESY OF JOHN MATHERLY/SHODAN (WWW.SHODAN.IO)

data. Like many a project that started out as an interesting hack, however, Shodan has since taken on some interesting, unexpected applications.

Over the past few years, Shodan users have uncovered a series of alarming network vulnerabilities in Internet-connected devices, including a nuclear reactor; the cyclotron at the Lawrence Livermore National Laboratory outside Berkeley, CA; a water treatment plant outside Houston; electric power generators; oil rigs; and even a crematorium.

Eireann Leverett, a researcher in the Centre for Risk Studies at the University of Cambridge, U.K., used Shodan to identify more than 100,000 vulnerable IoT devices in 2011, concluding these flaws left them vulnerable to attack by “malicious actors.” In a similar vein, Billy Rios at Google and Michael McCorkle of Boeing also have identified a series of serious security exposures across a wide range of connected industrial devices.

To Matherly’s surprise, many of these devices turned out to be special-purpose industrial computers: control systems that perform highly specific tasks, like regulating the flow of water and other utilities, transportation systems, and even entire power grids—all controlled over the network by remote supervisory staff.

Unlike the consumer-facing Websites that most of us can find readily using commercial search engines like Google, industrial control systems (ICS) have largely remained hidden in plain view, invisible to web crawlers. Since Shodan’s launch, however, it has shone an unforgiving light on some of these devices’ glaring security flaws.

“Industrial control systems have relied on security by obscurity,” says Mather, who now spends much of his time consulting with organizations on strengthening the network security of these devices.

Most of these devices rely on proprietary hardware and software protocols that tend to mask their vulnerabilities—but also make it difficult for security researchers to develop generalized and replicable approaches to security. “The more accessible the technology, the easier it is for people to find and fix vulnerabilities,” says Mather.

More troublingly, many vendors failed to treat these risks seriously, assuming these systems could only be addressed directly, rather than over

“It’s a technical problem, but it’s also closely tied to business interests ... these days, the way companies tend to look at security is as a loss leader.”

an external network. As a result, many hardware makers have tended to treat potential vulnerabilities lightly.

O’Brien feels these exposures stem not just from technical failures, but from a fundamental lack of industry focus on security. “It’s a technical problem, but it’s also closely tied to business interests,” he says. “These days, the way companies tend to look at security is as a loss leader.”

Moreover, customers for these systems—like, say, power plant operators—tend to resist adding layers of security, to ensure their ability to respond quickly in case of emergency. End-users within these organizations often see additional security controls—like layers of password prompts—as more of a burden than a benefit.

Given the lack of customer demand, product managers at hardware companies often find it difficult to justify investing resources in preventive security measures that do not add new functionality. Complicating matters further is the difficulty of sending updates and patches to these devices without user-initiated firmware updates—a common practice for Web-based software applications. As a result, these industrial devices can often remain vulnerable for extended periods of time.

“To be fair, many of these systems were designed before the age of ubiquitous connectivity,” says Mather, “so the engineers didn’t worry about hardening their device against software attacks.”

Mather also points to economic factors at play: “There wasn’t a push by the ICS operators to demand better computer security from the manufac-

turers.” Instead, they tended to focus more on issues of availability and reliability, and treated security as a secondary consideration.

That is now starting to change, thanks in part to the visibility that Shodan has brought to these vulnerabilities. In a similar vein, an open source project called Onionscan has made considerable headway in exposing the possible vulnerabilities of physical devices over the Internet.

Looking ahead, Shodan is focused on developing more sophisticated tools and visualizations to make the data more accessible to non-technical users.

Elsewhere, Nathan Freitas of the Guardian Project is spearheading an effort to use Tor—a free software package often used by hackers and journalists to protect their privacy via a worldwide network of volunteer-run servers—to safeguard IoT devices by means of Home Assistant, a Python-based system that allows for Tor to be used for physical devices. The system relies on a Raspberry Pi computer running Tor’s software to mask the location of smart home devices by means of an authenticated hidden service that prevents anyone from locating and connecting to the devices without access to a passcode that the developers describe as a “cookie.”

While this technology remains in the experimental stage, Freitas hopes it will pave the way for more fully developed commercial IoT security applications in the future.

Amid the rise of connected devices and growing public concern about Stuxnet-style attacks on major infrastructure projects, the conditions seem ripe for IoT security applications to find more traction in the marketplace. Yet Mather feels the industry at large remains too blithe about these dangers.

“We keep deploying new devices that are insecure-by-default,” he explains. “The vulnerable IoT devices of today that get installed are going to stick around a long time and they have access to the internal networks of many homes and businesses.”

That might seem like a borderline-paranoid fantasy, but the rapidly accelerating development of “smart hardware” devices may bring these risks closer to home—and soon. For example, some firms are developing light bulbs that serve as Internet

hubs, relaying Wi-Fi signals, connecting thermostats, or even interfacing with a home security system. As these everyday devices become increasingly interconnected, the security risks multiply exponentially.

O'Brien believes the long-term solution to IoT security will require a balanced approach to technological innovation and public policy-making to create a more reliable physical computing infrastructure.

"We're at a point where we're re-thinking what the role of government ought to be," he says. Despite a number of high-profile cyberattacks in recent years, securing critical infrastructure remains an area of unclear jurisdictional ownership within the federal government, partially involving the Federal Bureau of Investigation (FBI), Department of Homeland Security, and other agencies. The White House recently released a statement clarifying the role of first responders in cybersecurity attacks, but there is plenty of work left to do on this front.

Meanwhile, Mather worries people fail to recognize the growing so-

Securing critical infrastructure remains an area of unclear jurisdictional ownership within the U.S. federal government.

phistication of seemingly everyday devices like light bulbs or coffee machines that are fast becoming part of a deeply interconnected—and potentially insecure—worldwide network of smart objects. "Those devices are full-fledged computers nowadays," he explains, "and with the increasing number of IoT devices that are being deployed, those vulnerabilities become a real concern."

Further Reading

Leverett, E.

Quantitatively Assessing and Visualising Industrial System Attack Surfaces. University of Cambridge Computer Laboratory, Darwin College, June 2011. <http://www.cl.cam.ac.uk/~fms27/papers/2011-Leverett-industrial.pdf>

National Science and Technology Council, *Federal Cybersecurity Research and Development Strategic Plan: Ensuring Prosperity and National Security*, 2016. https://www.whitehouse.gov/sites/whitehouse.gov/files/documents/2016_Federal_Cybersecurity_Research_and_Development_Strategic_Plan.pdf

Olsen, M., Schneier, B., Zittrain, J.

Don't Panic: Making Progress on the 'Going Dark' Debate. Berkman Center for Internet and Society, Harvard University, February 1, 2016.

https://cyber.law.harvard.edu/pubrelease/dont-panic/Dont_Panic_Making_Progress_on_Going_Dark_Debate.pdf

Seitz, J.

Dark Web OSINT With Python and OnionScan. Automating OSINT, July 28, 2016.

<http://www.automatingosint.com/blog/2016/07/dark-web-osint-with-python-and-onionscan-part-one/>

Alex Wright is a writer and researcher based in Brooklyn, NY.

© 2017 ACM 0001-0782/17/1 \$15.00

Milestones

New CS Education Framework for U.S. Schools

ACM is part of a committee of computer science organizations that has released a framework to inform implementation of computer science education in K-12 schools throughout the U.S.

ACM, Code.org, the Computer Science Teachers Association, the Cyber Innovation Center, and the National Math and Science Initiative recently announced the launch of the K-12 Computer Science Framework, intended to inform the development of standards, curriculum, and computer science pathways, and also to help school systems build capacity for teaching computer science.

Developed through partnerships with states, districts, and the computer science education community, the K-12 Computer Science Framework is a significant milestone for computer science in the U.S. It promotes a vision in which all students critically

engage in computer science issues; approach problems in innovative ways; and create computational artifacts with a personal, practical, or community purpose.

The framework is not a set of standards; it is a set of guidelines put forth by the community that can inform standards, curricula, and many other supports for computer science education. The framework's learning progressions describe how students' conceptual understanding and practice of computer science grow more sophisticated over time. The concepts and practices are designed to be integrated to provide authentic, meaningful experiences for students engaging in computer science.

"The K-12 Computer Science Framework not only includes technical concepts about computing, but also stresses

the importance of creating an inclusive culture in the field, promoting collaboration among students, and communicating effectively about technology," said Mehran Sahami, Associate Chair for Education in the computer science department at Stanford University. "In this regard, the framework provides skills that generalize beyond computer science while also giving students an understanding of fundamental computing concepts that will serve them well in whatever career they choose to pursue." Sahami also co-chairs ACM's Education Board and Education Council.

ACM, CSTA, INFOSYS ANNOUNCE AWARDS FOR CS TEACHING EXCELLENCE
Infosys Foundation USA, ACM, and CSTA, the Computer Science Teachers Association, recently announced the launch of the Awards for Teaching Excellence

in Computer Science. Up to 10 awards of \$10,000 each will be awarded annually.

Funding for the awards is being provided by Infosys Foundation USA. Mark R. Nelson, CSTA's Executive Director, said Infosys "is sending a powerful message to these computing educators worldwide that what they are doing is indeed important."

"Great computer science education starts with great teachers," explains ACM President Vicki L. Hanson. "This new award reinforces our long-held goals of recognizing the contributions of computer science teachers and building a framework that supports their professional development."

Winners of the 2016 awards were announced in December (after press time). The prizes will be awarded at the 2017 CSTA Annual Conference in Baltimore in July.

Bias in Technology

As leading companies release troubling diversity statistics, experts search for solutions.

THE TECHNOLOGY WORLD has a diversity problem. A recent U.S. Equal Employment Opportunity Commission (EEOC) report found that the high-tech industry employed far fewer African-Americans, Hispanics, and women, relative to whites, Asian-Americans, and men. The difference is especially glaring in Silicon Valley. At Google and Facebook, African-Americans represent just 1% of the tech work force. The numbers are slightly higher at some other leading technology firms, but still are hardly reflective of society at large.

In academia, the figures are also discouraging. According to the 2015 Taulbee Survey, conducted by the Computing Research Association, African-Americans represented only 4.6% of the students awarded bachelor's degrees in computer science (CS). Women represented 15.7% of the surveyed population, but this is a significant decrease from 1984–1985, when the National Center for Education Statistics found that women made up 37% of CS undergraduates.

Today, experts in both the corporate and academic worlds are working to understand the root of the imbalance and searching for ways to expand the number of women and minorities in technology—and keep them there.

Diversity Is Good Business

The necessity of a more diverse technology workforce is a matter of social equality, but there are other compelling factors as well. When groups are underrepresented on product development teams, for example, the resulting technology can be biased. “I always ask the question: Is it possible for you to invent or create anything in the absence of who you are?” asks University of Florida computer scientist Juan Gilbert. “You might say, ‘I write algorithms; there’s no bias in that.’ But we’ve seen examples where algorithms do have bias.”



Organizations like Code.org are working to expand access to computer science and increase participation by women and underrepresented students of color.

In 2015, due to a quickly corrected flaw in its facial recognition software, Google Photos sorted images of one user and her African-American friend into an album labeled “Gorillas.” In another 2015 incident, a white male and his African-American friend recorded video of themselves testing the electronic soap dispensers at an Atlanta hotel because the devices did not register the presence of black skin.

Gilbert and others argue that a more diverse group of employees on the product development side could have prevented these kinds of problems. But there are also indications that broader representation can lead to increased revenue. Gilbert cites the story of the once-popular Motorola Razr mobile phone. When the product team was brainstorming potential colors for the phone, one of the women in the group suggested a more feminine hue for one of the models. She was laughed at initially, Gilbert says, but eventually the men in the group conceded. “They created a pink phone and it became the number one seller,” Gilbert says. “If she hadn’t been in the room, that wouldn’t have happened.”

Still, the business argument for inclusion may be simpler than that. “If you really want everyone in the world to use your products, then you need to have everyone working for you,” says Kaya Thomas, an African-American woman majoring in computer science at Dart-

mouth College. “If you want to sell to everybody, you have to hire everybody.”

The Pipeline Problem

Last July, when Facebook revealed it had made little to no improvement in its diversity numbers relative to the prior year, the company attributed the shortfall in part to a pipeline problem. The statement prompted significant criticism, with many citing Taulbee Survey data as evidence that qualified students are being overlooked: if 4.6% of CS graduates are African-American, why are only 1% or 2% of new tech-related hires at some major companies African-American? “The students are out there,” says Ann Quiroz Gates, chair of the Department of Computer Science at the University of Texas at El Paso. “I don’t agree the talent is not there. The talent is there.”

Gates says many of the skills companies look for in students can be acquired outside the classroom by competing in hackathons or building a GitHub profile. Unfortunately, her minority students often do not have the spare hours to participate in such extracurricular activities. “Finding the time to program beyond classroom assignments and experimenting with computers is really difficult because many of our students work full time while carrying a full course load,” she says. “They’re exhausted.”

The high-tech companies Gates has engaged with are starting to understand that, and she is working with several companies to set up paid positions at local firms, or university-funded projects that would allow her students to acquire those outside-the-classroom problem-solving skills and still earn the money they need to pay for their education.

While these efforts may help existing students land sought-after positions, the Taulbee Survey indicates more needs to be done to encourage women and minorities to pursue CS in the first place—and to stick with the subject. Maja Mataric, a computer scientist at the University of Southern California, began her career in the 1980s, when the participation of women in CS was on the upswing, so she has observed the decline firsthand. “We did really well for a while, but we never reached equality at all,” she notes.

In her view, several factors influenced the drop. The culture of male-dominated departments, for example, either ended up pushing women away or encouraging them to assimilate and sustain the very culture that was discouraging others. Mataric also points to how departments often promote women-in-computing organizations such as the Anita Borg Institute as support groups. This suggests the women selected to attend need this support, and are somehow weak. Instead, Mataric says, their selection should be presented as a sign of excellence.

These subtle cues add up, according to Mataric. “It is a very, very slanted, un-level playing field,” she says, “and the only people who can survive are either so driven that they don’t care or they’re just bullheaded. But that’s a small subset.”

Changing the academic culture can only solve part of the problem. Mataric believes the effort to increase the representation of women in technology needs to begin at an early age. In her own outreach projects, she focuses on introducing elementary school students to computational thinking through robotics. Part of her strategy is to show kids that software engineering is not just coding in the dark. “I really think the only way to fight it is from the very beginning, because so much of what we’re doing is just patching along the way.”

Kinnis Gosha, an assistant professor of computer science at Morehouse College, a private, all-male, liberal arts, historically black college in Atlanta, lauds projects such as Black Girls Code for spiking interest in the field among middle school students, but he says such programs are just the first step. Gosha, who is also the director of the college’s Culturally Relevant Computing Lab, calls for increased focus on improving high school-level computer science education. If college students want to compete for prestigious internships at the top technology companies and then land the best jobs, he argues, they need to start building their skills in high school by taking classes such as Advanced Placement (A.P.) Computer Science.

In his own courses at Morehouse, Gosha has seen firsthand that a student with this experience has a tremendous head start. “A student able to take that class and do well is a year or more ahead of the others,” he says. Unfortunately, the program is not always an option for kids. “In the Atlanta public school system,” Gosha notes, “there are only two high schools that offer the A.P. Computer Science course.”

If You See It, You Can Be It

Still, a student who starts late can enjoy a bright future. Kaya Thomas, the Dartmouth undergraduate, was not interested in computer science in high school. Then, during the winter break of her freshman year, Thomas started reading about the lack of women, and particularly women of color, in computer science. She took an online coding course during the break, fell in love with the work, and switched her major when she returned to school. Now, as a Fellow of CODE2040, the nonprofit that helps African-American and Latino computer science majors, she’s part of a community of woman and minorities like her, and she has already interned at three major technology companies.

She is the only African-American woman in her class of approximately 90 computer science majors at Dartmouth, but she is mentoring younger students, and encouraging them to stay in the field. “There are a lot of black girls—and by a lot I mean five or six—majoring in computer science in the sophomore class,” she says with

pride. But such small examples of success could continue to build on one another. “If you never see anyone who looks like you in a space, it’s a challenge for you to be in that space,” says Gilbert. “If you see it, you can be it.”

Of course, successful minority or female software engineers and developers cannot spend all their time mentoring. With that in mind, Gosha and several colleagues have been working on a virtual mentoring program that would feature interactive avatars of minority tech workers and executives responding to common questions. The avatars would not replace face-to-face mentoring; they would complement it, or even offer a young person their first exposure to a successful minority tech worker. “The idea is to virtualize the experience of meeting another black person who actually has a computing job, because African-American youths might not be in places where those people are visible,” he explains.

Overall, experts and insiders believe improving diversity in the technology world will require continued effort from academia, corporations, nonprofits, universities, and others, but Gilbert says the impact of these projects will reach beyond the business world. “We can’t be as great as we can be as a country unless we increase the participation of everyone,” he says. “Diverse minds will give us the best solutions to problems.” **C**

Further Reading

Diversity in High Tech, U.S. Equal Employment Opportunity Commission, 2015. <http://www.eeoc.gov/eeoc/statistics/reports/hightech>

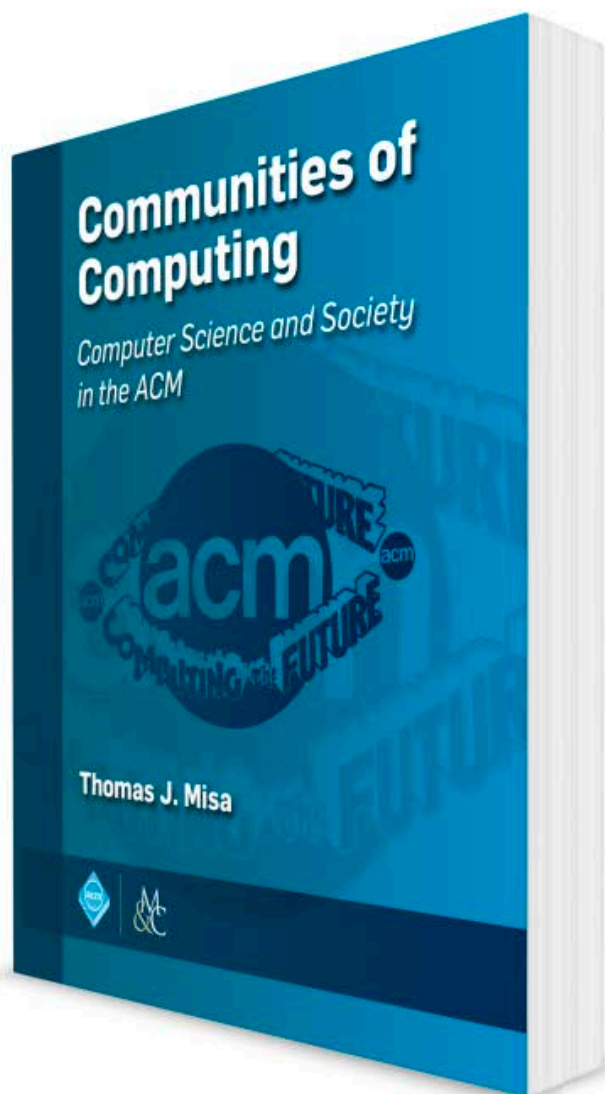
Hewlett, S.A., Luce, C.B., Servon, L.J., et. al. The Athena Factor: Reversing the Brain Drain in Science, Engineering, and Technology, a *Harvard Business Review* Research Report, 2008.

Intel Corporation & Dalberg Global Investment Advisors. Decoding Diversity: The Financial and Economic Returns to Diversity in Tech, 2016. <http://bit.ly/2bHEFuN>

Zweben, S. and Bizot, B. 2015 Taulbee Survey: Continued Booming Undergraduate CS Enrollment; Doctoral Degree Production Dips Slightly, *Computing Research News*, May 2016, Vol. 28 / No. 5

Gregory Mone is a Boston, MA-based writer and the author of the novel *Dangerous Waters*.

© 2017 ACM 0001-0782/17/1 \$15.00



**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>



DOI:10.1145/3018990

Michael A. Cusumano

Technology Strategy and Management

Is Google's Alphabet a Good Bet?

A relatively simple query raises myriad complicated issues.

GOOGLE INC., FOUNDED in 1997 by Larry Page and Sergey Brin, went public in 2004 to a great deal of fanfare (see “Google: What It Is and What It is Not,” *Communications*, February 2005). It reached another milestone in October 2015 when it reorganized as Alphabet Inc. The company was the second most valuable firm in the world as of November 2016, worth around \$528 billion, not far behind Apple (\$589 billion) and ahead of Microsoft (\$468 billion). The Google Internet business still centers on a unique search technology based on page-rank algorithms and generates enormous revenues from targeted advertisements and sponsored ads. Scale economies and network effects (that is, the more users of Google search, the more accurate the searches and ads become) also have contributed to Google's success. But is Google's transformation into Alphabet Inc. a good bet—for Google investors and users, and society more broadly? That

Google has always experimented with new product and service ideas as well as acquisitions, some more successful than others.

simple question raises big issues, such as how much should we expect large corporations to invest in research that might benefit society but not their bottom lines, and how might large corporations better use the money they do invest in research and new ventures?

It was never the case that Google's founders saw their future as limited to search. Google has always experimented with new product and service ideas as well as acquisitions, some more

successful than others. Google+, introduced in 2011 to compete with Facebook, has struggled. But Gmail, introduced between 2004 and 2007, Google Maps, introduced in 2005, and YouTube, acquired in 2006 (for \$1.65 billion), have greatly expanded Google's Internet platform. In 2005, Google really hit a homerun when it bought a small company called Android for an estimated \$50 million.⁹ In 2007, it used the technology to launch a new mobile operating system to compete with Apple's iOS. Google Android now leads the industry with over 80% market share.^a Largely on the strength of mobile searches from Android phones and tablets, Alphabet's revenues in 2015 were nearly \$75 billion, with operating profits of over \$19.3 billion—a profit rate of 26%, compared to 30% for Apple and 19% for Microsoft and IBM.

Let's look more closely at how and why Google reorganized. The new structure created a holding company that

^a See <http://bit.ly/2d7iCPb>.



breaks out the core and non-core operations (including research) into separate subsidiaries (internal divisions). This change clarifies which operations make money and which do not. The largest subsidiary retains the name Google. Now headed by Sundar Pichai, this division controls the related Internet businesses such as Search, Gmail, Android, Chrome, YouTube, Maps, Google Play, and Ads. Alphabet lumps the other subsidiaries into a single category called “Other Bets.” These consist of Nest, Google Access (formerly Google Fiber), Calico, Verily, Google Ventures, Google Capital, Jigsaw, Sidewalk Labs, and the Google X research labs.

Some Other Bets, such as Nest and Fiber, are real, albeit modest, businesses. Like Android, Google Maps, and YouTube, they are not all the results of in-house research projects. Google bought Nest Labs (<https://nest.com>) in 2014 for \$3.2 billion—now seen as a high price given current revenues estimated to be a few hundred million dollars per year.⁶ Nest sells smart ther-

mostats and smoke detectors, as well as indoor and outdoor webcams (from another acquisition). It is developing other devices for the “smart home” and the Internet of Things that may someday sell in volume. Google Fiber, started in 2010 as a research project, brings very-high broadband speeds to millions of subscribers. Now called Access, the company sells a 1Gbps connection for \$70 a month, the fastest in the industry, but still has relatively few subscribers due to limited availability. It has also incurred enormous capital expenses—\$6 billion to lay cables in just six cities. One estimate is that it will take 40 years to recoup the investment simply on the first six cities, yet Access has plans for 16 more cities. Not surprisingly, Alphabet CEO Larry Page recently ordered that employee numbers be cut in half and other expenses reduced.⁷

Other Bets subsidiaries are all early-stage ventures or research organizations, mostly intended to benefit society and not necessarily become great new businesses in the short term. So inves-

tors should not expect them to generate profits now—but what about in the future? After all, Alphabet is a for-profit company, supposedly focused on organizing the world’s information and then providing various Internet-based services. Some new ventures don’t fit this vision at all. For example, Google established Calico (<https://www.calicolabs.com/>) in 2013 to tackle aging and related diseases, with the primary goal of extending life. It has not announced any drugs or therapies, but has hired prominent scientists with backgrounds in cellular biology, genetics, synthetic biology, opto-genetics (using light to manipulate cells in living tissue), and cancer research.¹ By contrast, Verily (<https://www.verily.com/>) dates back to a Google X project launched in 2012 to put computing inside contact lenses. This effort led to the Verily spinoff in 2015. With about 400 employees in 2016, Verily has continued to work on contact lenses that can do tasks such as measure glucose levels for diabetics. Other Verily projects include a health-tracking

Google Alphabet's annual report differentiates core Google operations from Google's Other Bets subsidiary.

	Year Ended December 31		
	2013	2014	2015
REVENUES			
Google	\$55,507	\$65,674	\$74,541
Other Bets	12	327	448
TOTAL REVENUES	\$55,519	\$66,001	\$74,989
SEGMENT OPERATING INCOME (LOSS)			
Google	\$16,260	\$19,011	\$23,425
Other Bets	(527)	(1,942)	(3,567)
Reconciling Items*	(330)	(573)	(498)
TOTAL OPERATING INCOME	\$15,403	\$16,496	\$19,360
CAPITAL EXPENDITURES			
Google	\$7,006	\$11,173	\$8,849
Other Bets	187	501	869
Reconciling Items**	165	(715)	197
TOTAL CAPITAL EXPENDITURES***	\$7,358	\$10,959	\$9,915

* Reconciling items are primarily related to administrative costs not allocated to segments

** Reconciling items are primarily related to timing differences of payments as segment capital expenditures are on accrual basis while total capital expenditures shown on the Consolidated Statements of Cash Flow are on cash basis, capital expenditures of Motorola Mobile and Home, and other miscellaneous differences.

*** As presented in Consolidated Statements of Cash Flow.

wristband, surgical robots, and health-related “baseline” databases. The goal of transforming healthcare with technology remains core to the company and does fit, more or less, with Google’s original vision, though many key researchers have recently left due to vague objectives and leadership conflicts.¹⁰

Larry Page made it clear years ago that he intended to use Google’s profits to invest in “moonshots”—new product and service experiments that promised revolutionary innovations and “10x” performance breakthroughs.⁸ These goals motivated him to create Google X in 2010 as a research organization that has since worked on everything from self-driving cars and wearable computers to machine translation, artificial intelligence, and quantum computing (see <https://research.google.com/>). The labs sometimes launch initiatives that help the core businesses, such as machine translation or use of AI in search. But the stated mission is really to tackle bold new projects that potentially could change the lives of billions of people. Although many proposals seem to border on science fiction, to be approved, they must also put together technologies that already exist or nearly exist.⁵ The current

set of initiatives include the self-driving car, more development on Google Glass (the head-mounted display and wearable computer that, in beta release, had trouble generating third-party applications and user adoption), Loon (using balloons to bring Internet access to remote areas), and Wing (drones for package delivery, similar to the Amazon initiative).

In 2012, Google also restructured its shareholding system into dual share classes in order to allow the company founders to control voting rights.¹¹ So, in fact, Larry Page and other company leaders do not have to worry too much about shareholder pressures for short-term profits. Nonetheless, analysts and investors have criticized him for spending so much money on projects that do not help Google’s core businesses.⁴ The Alphabet reorganization does nothing to address these criticisms except to clarify where money is going—but only in the aggregate. For example, Alphabet’s 2015 *Form 10-K* annual report now breaks out the core Google operations from Other Bets (see the accompanying table).^b We

b Source: Alphabet Inc., *Form 10-K* (Washington D.C., United States Securities and Exchange Commission), December 31, 2015, p. 95 (Note 16)

can see that the Other Bets subsidiaries have grown revenues from \$12 million in 2013 to \$448 million in 2015. If we read the fine print in the annual report, we also learn that the revenue increase comes largely from the sale of Nest devices. At the same time, Other Bets’ operating losses grew from \$527 million in 2013 to a staggering \$3.6 billion in 2015.

Meanwhile, Alphabet increased overall R&D expenses in 2015 (these include Other Bets as well as Google) to over \$12 billion. This sum represented 16.3% of total revenues, up from the 12.8% of revenues that Google averaged from 2006 to 2013. Without the nearly \$3.6 billion loss from Other Bets, in 2015 Alphabet would have had an operating profit of about 36% of sales, rather than 26%. Maybe this money is being well spent—but maybe not. The latest Alphabet R&D expenditures are especially high compared to Apple, which, in 2015, spent only 3% of its revenues on R&D. Amazon, Microsoft, Cisco, and Oracle all invested between 12% and 14%, much closer to Alphabet. IBM spent 6%. Among other large technology firms, the leaders in R&D spending in 2015 were Intel (22%), Facebook (27%), and ARM (29%), though most of these expenditures appear directed

at incremental product development rather than moonshot-type research.

Perhaps the biggest implication from the Alphabet reorganization is what it says about the company's commitment to experimentation. There was a time when major corporations like AT&T (with Bell Labs) and Xerox (with Xerox PARC) made bold investments that generated whole new industries like semiconductors, personal computers, and networking technologies. These innovations were only tangentially related to their core businesses and mostly benefitted other firms (Apple, Microsoft, Intel, and other chip manufacturers), as well as society. As a result, most big corporations over the past two decades have scaled back or even disbanded basic research that did not tie in to product divisions and the large, centralized research organizations needed to sustain these efforts.² Instead, we find research universities taking the lead in searching for scientific breakthroughs and some new applications intended to benefit society, though often the universities work in partnerships with corporations, government labs and programs, and some venture funds.

So is Google and now Alphabet a good bet for investors given its commitment to "moonshot" research, related and unrelated to its core businesses? Most projects have not produced much revenue so far, and they are a growing drag on profits.³ Maybe this would not be an issue if a corporation has enough resources (people and money) to do both long-term research and to keep improving its main products and services. However, we have not seen many breakthroughs from Google X in search technology, operating system design, application development environments, or related Internet services.

It is worth noting that other technology billionaires, such as Microsoft founders Bill Gates and Paul Allen, as well as Amazon founder Jeff Bezos, also have invested vast sums in bold initiatives unrelated to the businesses of their original companies. But they have done so mostly in separate ventures and mostly with personal funds, including donations to foundations, and other investor money. Elon Musk, with Tesla, SpaceX, and Solar City, is also a big personal investor in

moonshot-type ventures, but he is also using funds from investors who should know what they are doing.

Here the transparency we see in the Alphabet restructuring is an improvement. As long as Alphabet-Google has the money, and investors know what they are investing in, it is important for society that someone takes a chance on risky new ventures. We will never reach the moon again or make equivalent leaps forward in science and technology if no one tries. AT&T Bell Labs is not what it once was and Xerox PARC no longer exists. IBM, Microsoft, General Electric, and a few other companies continue to do some basic research, but they are all under pressure to assist their product divisions and generate new billion-dollar businesses.

Alphabet's Other Bets may yet produce some great new businesses for the company as well as contribute to the betterment of humankind. In the meantime, though, shareholders need to understand that the money is coming out of their pockets, and so it would be useful for Alphabet to be even more transparent with details on the individual bets. In addition, Larry Page and other company leaders should acknowledge more openly that these expenditures may also be at the expense of innovations that could more directly benefit the billions of current Google users as well as ensure that the company remains healthy long into the future. **C**

References

1. Bergen, M. What's he building in there? The stealth attempt to defeat aging at Google's Calico. *Recode* (Dec. 28, 2015).
2. Chesbrough, H. Is the central R&D lab obsolete?" *Technology Review* (Apr. 24, 2001).
3. Fiegerman, S. Google's moonshot projects lost \$859 million last quarter. *CNN Money* (July 28, 2016).
4. Gaudin, S. Google's moonshots increasingly expensive. *Computerworld* (Apr. 22, 2016).
5. Gertner, J. The truth about Google X: An exclusive look behind the secretive lab's closed doors. *Fast Company* (Apr. 15, 2014).
6. Hesler, Y. Even Google views its Nest acquisition as a disappointment. *BGR* (Mar. 31, 2016).
7. Levy, A. Google Fiber has been a huge disappointment. *The Motley Fool*, (Aug. 29, 2016).
8. Levy, S. Google's Larry Page on why moonshots matter. *Wired* (Jan. 17, 2013).
9. Melanson, D. Google execs call Android acquisition its best deal ever. *Engadget* (Oct. 27, 2010).
10. Piller, C. Google's bold bid to transform medicine hits turbulence under a divisive CEO. *STAT* (Mar. 28, 2016).
11. Solomon, S.D. New share class gives Google founders tighter control. *The New York Times* (Apr. 13, 2012).

Michael A. Cusumano (cusumano@mit.edu) is a vice president and dean at Tokyo University of Science, on leave as a professor at the MIT Sloan School of Management.

Copyright held by author.

Calendar of Events

January

January 18–20

FOGA '17: Foundations of Genetic Algorithms XIV, Paris, France, Sponsored: ACM/SIG, Contact: Giuseppe Castagna, Email: Giuseppe.Castagna@cnrs.fr

February

February 4–8

PPoPP '17: 21st ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, Austin, TX, Sponsored: ACM/SIG, Contact: Vivek Sarkar, Email: vsarkar@rice.edu

February 6–10

WSDM 2017: Tenth ACM International Conference on Web Search and Data Mining, Cambridge, U.K. Co-Sponsored: ACM/SIG, Contact: Milad Shokouhi, Email: milads@microsoft.com

February 25–Mar 1

CSCW '17: Computer Supported Cooperative Work and Social Computing, Portland, OR, Sponsored: ACM/SIG, Contact: Steven E. Poltrock, Email: spoltrock@gmail.com

February 26–28

FPGA '17: The 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Monterey, CA, Sponsored: ACM/SIG, Contact: Jonathan Greene, Email: jonathan.greene@microsemi.com

March

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

Law and Technology

Why Less Is More When It Comes to Internet Jurisdiction

Considering legal uncertainty in the online environment.

IN THE DAYS before widespread broadband, social networks, and online video, a French anti-racism group launched the Internet lawsuit heard round the world. In late 1999, the International League against Racism and Anti-Semitism—or Ligue Internationale Contre le Racisme et l'Antisémitisme (LICRA) in French—filed suit against then-Internet giant Yahoo, seeking a court order to compel the company to block French residents' access to postings displaying Nazi memorabilia. While Yahoo already blocked access to content on its local French site (<http://www.yahoo.fr>), the lawsuit targeted the company's primary site based in the U.S. (<http://www.yahoo.com>).

The case attracted immediate interest since it struck at the heart of one of the Internet's most challenging issues—how to bring the seemingly borderless Internet to a bordered world. Given that the Internet has little regard for conventional borders, the question of whose law applies, which court gets to apply it, and how it can be enforced is seemingly always a challenge.

Striking the right balance can be exceptionally difficult: if courts are unable to assert jurisdiction, the Internet becomes a proverbial “wild west” with no applicable law. Conversely, if every court asserts jurisdiction, the Internet becomes overregulated with a myriad of potentially conflicting laws vying to govern online activities.



The temptation for politicians, courts, and regulators is invariably to assert jurisdiction over online activities regardless of the impact on other countries or the potential conflict with different rules. Yet experience suggests that when it comes to Internet jurisdiction, less is often more. While few still argue the law does not apply online, exercising restraint in asserting jurisdiction is likely to increase global respect

for the law and better ensure enforcement of judicial decisions.

Where It All Began: The Yahoo France Case

The Yahoo France case sparked nearly six years of litigation, numerous legal briefs, and much hand-wringing from the Internet community. The initial French court ruling sided with LICRA and ordered Yahoo to do what it reason-

ably could to ensure that French users could not access content that was unlawful in France.^a The judge was persuaded at that time that Yahoo was capable of identifying when French users accessed its site (it provided targeted ads to such users) and that his order would be limited to Yahoo's activities in France.

Yahoo was unsurprisingly critical of the decision, but rather than appealing the French decision, it chose to let the decision stand and to launch a lawsuit of its own in the U.S. courts, seeking an order that the French decision could not be enforced on its home turf.

The 9th Circuit Court of Appeals, a U.S. appellate court, ultimately issued a 99-page split decision that asserted jurisdiction over the dispute but declined to provide Yahoo with its much-desired order.^b The U.S. decision turned on the fact that Yahoo had independently removed much of the offending content, suggesting that the company was not being forced to block legal materials.

On the question of jurisdiction, the majority of the court determined that it could assert jurisdiction over the case despite minimal connections to the U.S. Indeed, in this case the contacts were limited to a cease and desist letter demanding that Yahoo comply with French law, the formal delivery of the lawsuit, and the mere existence of the French court order.

The French and U.S. courts both demonstrated that the default in most Internet jurisdiction cases is to assert jurisdiction, even if doing so is likely to lead to conflicting decisions, thorny conflict of law issues, and regulatory uncertainty.

Is Internet Publication Anywhere or Everywhere?

The Yahoo France case may have been the first major Internet jurisdiction case to attract global attention but it was by no means the last. Several years after the Yahoo decision, the *Washington Post* found itself at the center of a Canadian case that combined Internet

The broader implications of the ruling strike a chord with those concerned with legal overreach on the Internet.

jurisdiction principles with the global availability of media publications.

The case concerned a defamation suit launched against the *Post* by Cheikh Bangoura, a former U.N. official. Bangoura had moved to the Province of Ontario several years prior to the lawsuit, but was stationed in Kenya in 1997 in a U.N. Drug Control Program when the *Washington Post* featured several articles accusing him of misconduct and mismanagement.

Bangoura sued the *Washington Post* in the Ontario courts in 2003, claiming the articles were untrue yet remained available on the *Washington Post* website and therefore accessible to residents in Ontario. The newspaper sought to have the case dismissed, arguing the Ontario courts should not assert jurisdiction over the matter since there was no real and substantial connection with the province.

In a surprise decision, an Ontario judge denied the *Washington Post*'s motion, ruling the paper "should have reasonably foreseen that the story would follow the plaintiff wherever he resided."^c

Dozens of global media organizations banded together to support the *Washington Post* in its appeal. The Ontario Court of Appeal sided with the newspaper, noting that "the connection between Ontario and Mr. Bangoura's claim is minimal at best. In fact, there was no connection with Ontario until more than three years after the publication of the articles in question."^d

Given that analysis, the court concluded that "it was not reasonably foreseeable in January 1997 that Mr. Bangoura would end up as a resident of Ontario three years later. To hold otherwise would mean that a defendant could be sued almost anywhere in the world based upon where a plaintiff may decide to establish his or her residence long after the publication of the defamation."

The New Internet Jurisdiction Storm: Google

The Yahoo France and Bangoura cases provided early hints at the challenge of reconciling easy access of content from global media companies through the Internet with differing legal standards. Those cases foreshadowed an even larger legal battle: the ability for a single country or jurisdiction to limit global access to search results from Google.

Among the recent cases, perhaps the best known involves European privacy law. In 2014, the European Court of Justice ruled on the "right to be forgotten," which requires Google to remove links from its search index to certain content. The decision is grounded in European privacy law and was initially limited in application to the European Union. Yet in recent months, the global impact of the decision has become increasingly apparent.

The decision arises from a 2010 complaint by a Spanish man who was upset to find that searching his name in Google yielded links to a 1998 announcement in a newspaper on a real estate auction designed to generate proceeds to pay back social security debts. The information was both factual and readily accessible online, yet the man felt the now-outdated information was a violation of his privacy.

The court ruled that it could assert jurisdiction over the search giant, despite the fact that the processing of the data took place outside of Spain.^e In addition to asserting jurisdiction over Google, privacy authorities expanded the scope of the ruling by demanding the removal of links from all Google search indices, not just those in Europe. While Google initially resisted, the company acquiesced in

a *UEJF and Licra v. Yahoo! Inc. and Yahoo France*. Tribunal de Grande Instance de Paris (May 22, 2000); <http://bit.ly/2fjHokU>

b *Yahoo! Inc. a Delaware corporation v. La Ligue Contre Le Racisme et L'antisemitisme*. L'Union Des Etudiants Juifs De France, 433 F.3d 1199 (9th Cir. 2006); <http://bit.ly/2f8Oi59>

c *Bangoura v. Washington Post*, 2004 CanLII 26633 (ON SC); <http://bit.ly/2flqnT>

d *Bangoura v. Washington Post*, 2005 CanLII 32906 (ON CA); <http://bit.ly/2gl0K3v>

e *Google Spain SL, Google Inc. v. Agencia Española de Protección de Datos*, Case C131/12; <http://bit.ly/2g0oeC>

COMMUNICATIONS APPS

Access the latest issue, past issues, **BLOG@CACM**, News, and more.



Available for iPad, iPhone, and Android



Available for iOS, Android, and Windows

<http://cacm.acm.org/about-communications/mobile-apps>



Association for Computing Machinery

early 2016, announcing that it would block search results from all Google domains where the search originated in Europe. The extension of the ruling marks an expanded Internet jurisdiction approach, with the effective application of European law to search results around the world.

A similar question sits at the heart of a Canadian case involving Google that was heard by the Supreme Court of Canada in December 2016. The case started in the Province of British Columbia, where courts were asked to consider whether they could assert jurisdiction over Google and how far to extend an order to remove links from its search index. The Canadian court orders have thus far intentionally targeted the entire Google database, requiring the company to ensure that no one, anywhere in the world, can see the search results.^f

The broader implications of the ruling strike a chord with those concerned with legal overreach on the Internet since if a Canadian court has the power to limit access to information for the globe, presumably other courts do as well. While the Canadian courts did not grapple with this possibility, what happens if a Russian court orders Google to remove gay and lesbian sites from its database? Or if a Saudi Arabian court orders it remove Israeli sites from the index? The possibilities are endless since local rules of freedom of expression often differ from country to country.

In fact, the lower court rulings provided the sense that the courts felt that their reach needed to match Google's global footprint. While there is much to be said for asserting jurisdiction over Google—if it does business in Canada, then Canadian law should apply—attempts to extend blocking orders to a global audience could lead to a run on court orders that target the company's global search results.

That would leave two possible problematic outcomes: Google would selectively decide which court orders it wishes to follow or local courts would begin deciding what the rest of the world can access online. Either way, the overreach of the courts could lead

to legal conflicts online and potential suppression of freedom of speech on the Internet.

In Search of an Internet Jurisdiction Solution

The Internet is often characterized as a “wild west” where laws cannot be easily applied. Yet the danger of extra-territorial application of court decisions such as those involving Google is that it encourages disregard for the rule of law online, placing Internet companies in the unenviable position of choosing the laws and court orders they wish to follow. Moreover, if courts or companies openly disregard foreign court orders, legal certainty in the online environment is undermined.

Years of litigation starting with the Yahoo France case suggest there are no easy answers. However, any solution likely lies in developing standards that encourage comity and mutual respect for the applicability of the law on the Internet by ensuring that national sovereignty is respected. Indeed, a recent 2nd Circuit Court of Appeals decision denying a government demand that Microsoft disclose email messages and private information hosted on a server in Ireland points to the fact that courts may be awakening to the need to establish limits on the jurisdictional scope of their orders.

The emerging approach suggests that courts should only issue orders with substantial extra-territorial effect where it is clear that the underlying right and remedy are also available in affected foreign countries. Global takedown orders or decisions with substantial impact in other jurisdictions is likely to enhance the perception of the Internet as a wild west where disregard for the law is common.

For that reason, where there is uncertainty about the legal rights in other jurisdictions, courts should exercise restraint, recognizing that less may be more. Indeed, respect for the law online may depend as much on when not to apply it as do efforts to extend the reach of courts and court orders to a global Internet community. 

Michael Geist (mgeist@uottawa.ca) holds the Canada Research Chair in Internet and E-commerce Law at the University of Ottawa, Faculty of Law.

^f *Equustek Solutions Inc. v. Google*, 2015 BCCA 265; <http://bit.ly/1FiP15K>

Historical Reflections

Colossal Genius: Tutte, Flowers, and a Bad Imitation of Turing

Reflections on pioneering code-breaking efforts.

MAY 14, 2017, will be the 100th anniversary of the birth of someone you might not have heard of: William Thomas (“Bill”) Tutte. During the Second World War he made several crucial contributions to decrypting the Lorenz cipher used to protect the Nazi high command’s most crucial radio communications. This work provided the statistical method implemented electronically by Tommy Flowers, a telecommunications engineer, in the Colossus machines, which pioneered many of the electronic engineering techniques later used to build digital computers and network equipment.^a

Myths of Bletchley Park

The British code-breaking effort of the Second World War, formerly secret, is now one of the most celebrated aspects of modern British history, an inspiring story in which a free society mobilized its intellectual resources against a terrible enemy. That’s a powerful source of nostalgic pride for a country whose national identity and relationship with its neighbors are increasingly uncertain. Tutte’s centennial gives a chance to consider the broader history of Bletchley Park, where the codebreakers worked,



A scene from the 2014 film *The Imitation Game* with actor Benedict Cumberbatch portraying Alan Turing.

and the way in which it has been remembered. Some kinds of people, and work, have become famous and others have not.

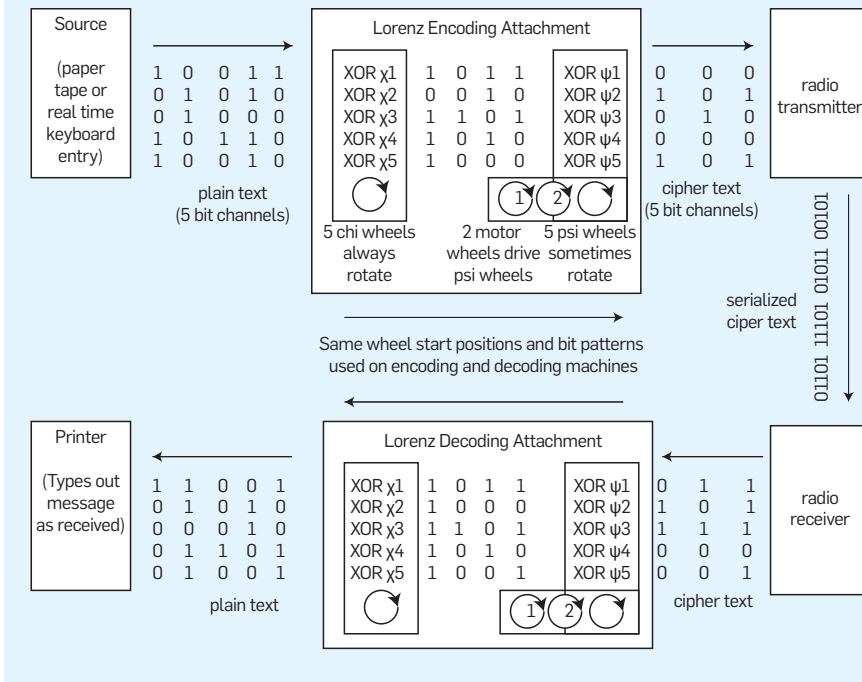
Films reach more people than books. So statistically speaking, most of what you know about Bletchley Park prob-

ably comes from the Oscar-winning film *The Imitation Game*. This gives us a starting point: the film is a bad guide to reality but a useful summary of everything that the popular imagination gets wrong about Bletchley Park.

One myth is that Alan Turing won

^a This column draws on a research project in progress I am conducting with Mark Priestley, supported by L.D. Rope’s Second Charitable Settlement.

Figure 1. Logical representation of the action of the Lorenz machine, dubbed “Tunny” by the British. The gap of four characters shown between the chi and psi wheels is to symbolize the idea of two logically independent transformations applied to a five channel bitstream, rather than a representation of the actual inner working of the machine.



the war pretty much by himself. Benedict Cumberbatch played Turing as television's *Sherlock*: a humorless, asexual loner whose superhuman mental powers are compromised only by an almost autistic indifference to social norms. This combines the traditional focus of popular science writing on the lone genius who changes the world with the modern movie superhero narrative of a freak who must overcome his own flaws before he can save the world. (This is, I understand, exactly the same arc Cumberbatch follows in *Doctor Strange*, a literal superhero movie). A team of six bright young men arrive at Bletchley Park and are given the job of breaking the German Enigma code. After introducing himself as “the best mathematician in the world,” Cumberbatch dismisses his fellow codebreakers, saying “I don't have time to explain myself as I go along, and I'm afraid these men would only slow me down.” Over their objections, and the opposition of almost everyone else at Bletchley Park, he designs a machine called Christopher to beat Enigma.

The film struggles to shoehorn a complicated, nuanced, and tragic story into the generic template of a hero overcoming obstacles, facing disaster,

and eventually triumphing. In reality, Turing was far from a marginal and despised figure at Bletchley Park, moving quickly to a leadership role. The obstacles Cumberbatch faces are mostly fictional and often absurd. His boss (in reality Turing's subordinate) Hugh Alexander tries to smash Christopher. Like all military men Alastair Denniston, the head of Bletchley Park, is a belligerent idiot. He smashes down the door, hoping to have Christopher destroyed. A little later Cumberbatch is blackmailed into passing secrets to a Soviet spy. Cumberbatch's growth is symbolized by his making friends with the other brilliant codebreakers, but they seem just as arrogant as he was. Fearful that Bletchley Park managers will get in the way, they pretend that Enigma is unbroken and instead run the Battle of the Atlantic from their machine room, deciding which convoys to save and which the U-boats should be allowed to sink.

Another myth is that code-breaking machines eliminated human labor and code-breaking skill. At first Christopher spins uselessly, producing dramatic tension but no Enigma settings. At the last minute, Cumberbatch thinks of looking for the common phrases, such as “Heil Hitler,” to detect

when the correct settings have been found. Enigma is broken! In reality this was not something that occurred to Turing only after he had finished building a completely pointless machine, but basic cryptanalytic practice already used to break Enigma manually.

Even with the Bombe, breaking Enigma continued to rely on a mix of human skill, manual methods, and machine power. Each Bombe was tended by a team of female operators, and in recent years a lot of attention has been paid to the experiences of these women. The Bombe operators and Enigma decrypters worked in factory-like conditions, in a massive operation pioneered by Gordon Welchman for the group attacking Air Force Enigma. Welchman had made his own vital contribution to the Bombe's design, a “diagonal board,” but after the machines arrived shifted to overseeing their use. In contrast, once logistical challenges began to outweigh conceptual ones Turing drifted away from leadership of the corresponding group tackling Naval Enigma. There are many kinds of genius, and Turing's was not of the managerial variety.

By early 1945 more than 10,000 people worked directly on the British code-breaking effort. Most of them were women. The film directs some heavy-handed sexism at Joan Clarke, a rare female cryptanalyst, to give viewers a chance to feel superior to their grandparents. Word at Bletchley Park was indeed deeply stratified by class and gender, and cryptanalytic work of the kind done primarily by upper middle class men like Turing has traditionally been the most celebrated. Yet its own erasure of the work routinely done by women reflects a more modern sexism. A single giant machine, given by a superman, clunks without human intervention until the answer emerges.

That's silly. So is the myth that brilliant scientists need no help from engineers. Cumberbatch drew up the engineering blueprints for the one and only Bombe during a montage, then built it at Bletchley Park aided by a handful of briefly glimpsed assistants. In fact, approximately 200 Bombes were manufactured, none of them at Bletchley Park, with design and production work done primarily by the British Tabulating Machine Company, under the direction of

its chief engineer Harold Keen. BTM had licensed IBM punched card technology for sale throughout the British Empire, giving it command of the necessary technologies and engineering methods.

In this way, and countless others, technology transcended, rather than supplemented, human labor and bureaucracy. Perhaps nobody would pay to see a movie centered on heroic production engineers, brilliant Bombe operators, and inspirational government procurement specialists, but would it be so difficult to give walk-on parts to the people who actually designed, built, and operated the Bombes? It's not like leaving them out makes the film particularly entertaining: political comedian John Oliver recently likened a hellish commute requiring three buses and two trains to "the length of the movie *The Imitation Game*—two wasted hours that feel like forever."

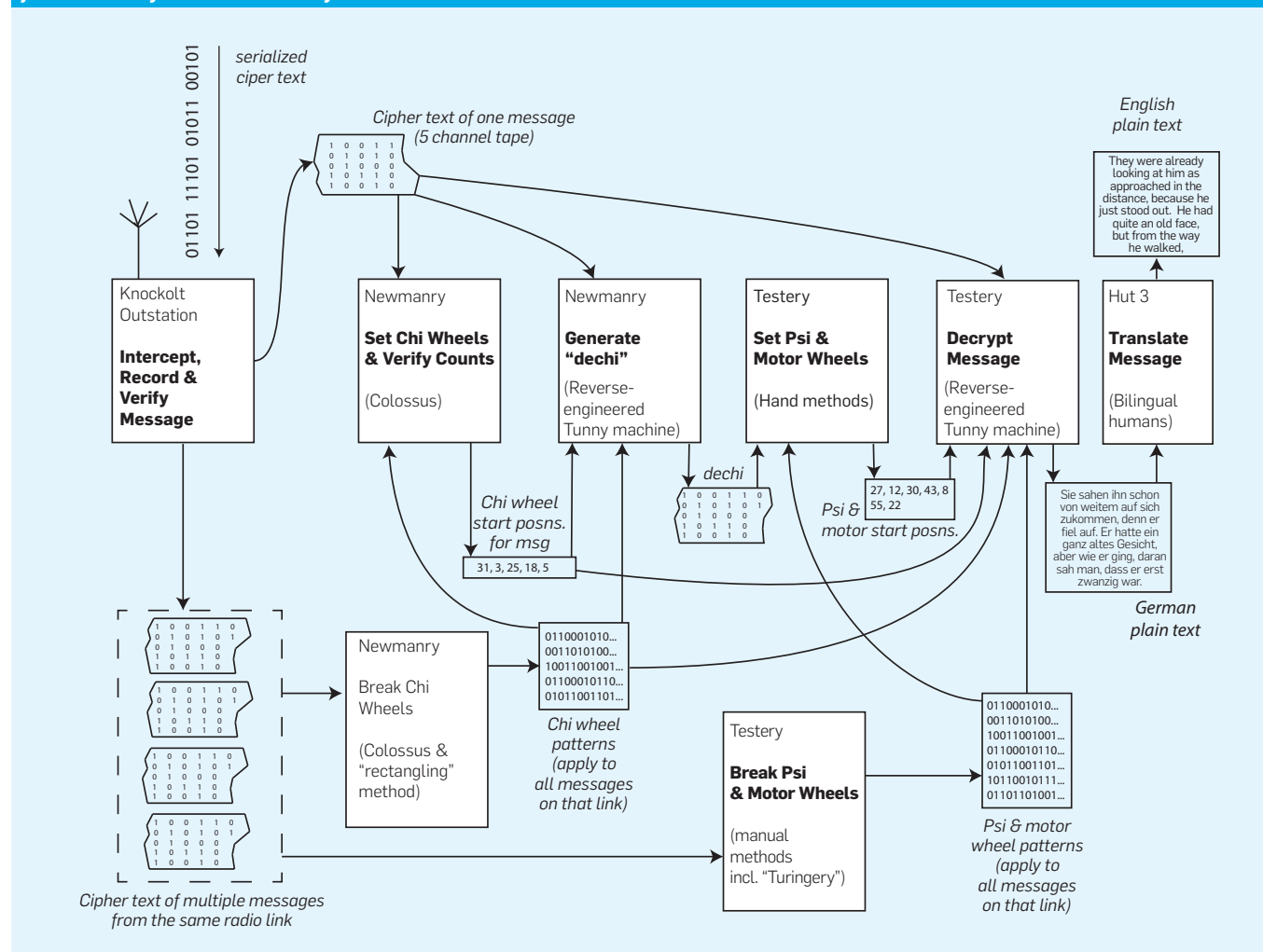
Tunny

There's another popular myth endorsed by the *The Imitation Game*: that the Enigma was the only significant German cipher broken at Bletchley Park. In reality another cipher, code-named Tunny, was just as important. Enigma was used by the German military to communicate with forces in the field. Compact code boxes flashed up the appropriate letter of cipher-text each time a key was pushed. The encrypted version of the message was transmitted by a human using Morse code. Tunny was used by the German High Command to communicate with generals outside the reach of wired communication links, such as the divisions deep inside the USSR. Messages typed on a keyboard were encrypted, transmitted by radio, and automatically decrypted and printed at the other end. Enigma was declassified long

before Tunny, with the result that early histories assumed that all intelligence delivered by Bletchley Park had come from Enigma.

The British knew nothing about the machine producing Tunny, working entirely on clues from the intercepted messages. In contrast, the team breaking Enigma had a number of resources to draw on. Enigma technology had been sold commercially since the 1920s, so its architecture was public. The German government specified modifications but Polish teams had broken earlier military versions. They passed their secrets, and the initial Bombe design, on to the British at the start of the war. Even the name "Bombe" came from the Polish "bomba kryptologiczna." Captured German personnel, code books, keys, and manuals provided more clues to changes in Enigma technology and practice.

Figure 2. The overall Tunny interception and decryption workflow process as of late 1944. While Colossus machines could also be used to tackle the Psi wheels, the increased pressure of work after the Germans began daily changes to wheel patterns meant that this part of the job was usually done in the Testery without Colossus assistance.



The first challenge was to figure out the structure of the machine that had produced the Tunny messages. This relied on operator error. In August 1941 a long message was sent twice, with minor alterations, using the same machine settings. Comparing the two messages let Brigadier John Tiltman piece together the original (“plain”) text of each message and the encoding sequence (“key stream”) with which each had been combined to yield its cipher text (see Figure 2). This nicely disproves the idea that Bletchley Park’s experienced military men were there primarily to frustrate Turing.

Tutte’s Breakthroughs

Tutte used these sequences to make the breakthrough. Shortly after the outbreak of war he had joined Bletchley Park’s “Research Section,” which had responsibility for investigating unfamiliar codes and devising methods that could be used to break them on a production basis. Like many of his colleagues he had a Cambridge degree, though as the son of a gardener his social background was unusually humble. His first degrees were in physical chemistry, but his interests were already shifting to mathematics.

Looking at the teleprinter tape five channels separately Tutte noticed a regularity every 41 characters in the key sequence. This gave the clue that a 41-character code wheel was involved. Step by step, Tutte’s team worked out the full, complex structure of the Lorenz machine and its 12 code wheels (see Figure 1). Three kinds of information were needed to decrypt a Tunny message. Tutte had provided the first, by reverse engineering the Tunny machine’s internal structure. From this the British built their own Tunny decrypting machines.

Two challenges remained. The second piece of information was the position of each of the wheels at the start of a message. Finding this was called “wheel setting,” and had to be done for each message. The third was the bit pattern set up on each of the wheels. These were regularly changed by flipping tiny pins. Finding these sequences was called “wheel breaking,” and was performed less frequently as most changed only monthly or quarterly.

After Tutte cracked the structure of

the Tunny machine a new group, named the Testery after its head, Major Ralph Tester, went to work breaking its messages. While seconded to the research section, Turing came up with a method for figuring out the bit patterns on the coding wheels once the keystream had been successfully isolated. The cryptanalyst formed and tested hypotheses to gradually extend the known portion of the wheel pattern. “Turingery,” described by Tutte as “more artistic than mathematical” made wheel breaking feasible, if arduous.

Both wheel setting and wheel breaking depended on luck, ingenuity, and German errors, but because wheel setting had to be done for each message it proved the biggest constraint. Until October 1942 the German operators were kind enough to precede each message with a series of 12 unencrypted letters, each coding the position for one of the wheels. Clues of this kind made the work of the codebreakers much easier, as hints obtained from decrypted messages were applied to others sent on the same link. Their results were sporadically impressive, but improvements in German practice threatened to close the door entirely.

At this point Tutte made his second huge contribution. When setting wheels it was easy to determine statistically whether a particular combination yielded natural language or random noise. Trying every possible combination of positions across the 12 wheels was clearly impossible: the war would be over, and the Earth swallowed up by the sun, long before the job was finished. But flaws in the design of the Lorenz machine made it possible to break the job into tractable steps. Each channel was encrypted by the successive action of two code wheels. The

Another myth is that code-breaking machines eliminated human labor and code-breaking skill.

first set, known at Bletchley Park as the chi wheels, rotated to the next position each time a character was read, whereas the second set, the psi wheels, turned only when directed to do so by two “motor wheels.” Decrypting a Tunny message posed two main challenges. First to set the chi wheels, using this information to generate “dechi,” text encrypted only by the psi wheels. Then to set the psi and motor wheels, using this information to generate plain text.

Because the psi wheels did not always rotate, their contribution to the cipher text often repeated from one character to the next. This, Tutte realized, gave a statistical method to set the chi wheels without making any assumptions about the psi wheels. Whether the wheels moved or not they still masked the distinctive character distributions of German text. But whenever the psi wheels did not rotate, the deltas between successive characters would pass through them unchanged.

Analyzing a sample of decrypted messages showed him that the distribution of deltas was far from random. For example, in German “ei” is a very common string. E is (1,0,0,0,0) and I is (0,1,1,0,0). Their delta, (1,1,1,0,0) had a frequency of 5.9%—almost twice as common as in a random distribution. The delta between two repeated characters, (0,0,0,0,0), occurred 4.6% of the time. German has many double “s” characters, and teleprinter operators often pushed the shift and unshift keys, encoded as their own characters, twice to make sure that they were received.

When the distribution of deltas in successfully dechided text was plotted the same tell-tale “bulges” in the distributions of deltas appeared. If the psi wheels moved about half the time the peaks and valleys would be half as tall but follow the same pattern. However, if dechi was produced with the wrong wheel settings the distribution of deltas should be close to random, with all combinations occurring about 3.1% of the time.

Setting chi wheels meant generating dechi with different wheel combinations and looking for a non-random distribution of deltas. The five chi wheels could take 22 million combinations, but because each acted on only one of the five bit channels that encoded each character there was no need to consider all wheels simulta-

neously. The five most common deltas were (1,1,0,1,1), (1,1,1,0,0), (0,0,0,1,0), (1,1,1,1,1), and (0,0,0,0,0). In each case first two channels were either (1,1) or (0,0) and so added to 0. Maybe your school didn't teach you that $1+1=0$, but in this context "addition" meant the logical operator XOR. Tutte devised a very simple method: (a) use all 1,271 possible wheel start positions to generate dechi for channels 1 and 2, (b) for each dechi stream count the number of positions where deltas for channels 1 and 2 add to 0, (c) take the wheel settings with the highest count. Once settings for the first two wheels were found the process was repeated to identify the others. If everything went well, this would set all five wheels by processing the encrypted message about 2,400 times.

Here, at last, was a way of breaking Tunny messages that did not depend on German laxity. It was possible to verify Tutte's method with hand calculation, but no manual method that involved comparing lengthy sequences so many times would ever be practical for production use. The success of the Bombes against Enigma made mechanization the obvious solution. Max Newman, a Cambridge mathematics lecturer, was put in charge of a small group exploring machine methods for Tunny.

Flowers and Colossus

Enter our third major protagonist, Tommy Flowers. Flowers came from a working class background, earning a degree in night school while working as an apprentice. He rose through the engineering ranks of the General Post Office, the government monopoly responsible for telephone and telegraph service as well as mail delivery. By the time the war began he was the leading electronics specialist at the Post Office's Dollis Hill research station. In January or February 1943, Newman arranged for a contract with Dollis Hill to build parts of a prototype machine able to apply Tutte's statistical method. Flowers, and others at Dollis Hill, had already been working with Bletchley Park to explore the application of electronic, rather than mechanical, sensing units to the next generation of Bombes.

Nicknamed "Heath Robinson," after a British cartoonist known for drawing improbably complex ma-

Technology transcended, rather than supplemented, human labor and bureaucracy.

chines, the prototype read two looped paper tapes. When implementing Tutte's method one would hold the message text and one the coding key series produced by a particular pair of chi wheels. Dollis Hill's experience with telegraphic equipment, and Flowers' personal enthusiasm for electronic signal processing, came together in an optical sensing system, able to coordinate the two tapes at up to 2,000 characters per second. This allowed it to evaluate all possible combinations for the first two chi wheels in about half an hour. After the prototype proved itself, during testing in mid-1943, Dollis Hill built several more Robinsons.

Back in March 1943, Flowers had also persuaded Newman to endorse experimental work on a machine able to simulate Tunny code wheels electronically rather than read them from tape. With no need to synchronize two tapes the message could be run at 5,000 characters per second. More importantly, it became much easier to change bit patterns or set new combinations of wheels. This reduced set up time between runs, and opened the door to new analytical tactics. As well as cramming electronic vacuum tubes into the simulated code wheels, Flowers used them liberally in its counters and in the logic circuits used to combine inputs in different ways. The first Colossus, delivered to Bletchley Park in January 1944, had approximately 1,500 tubes, a huge number by the standards of the day, but Flowers later said he knew it could run reliably because of his pre-war experiments with a prototype electronic telephone exchange.

Colossus delivered a huge productivity boost. In February Newman's section was setting only four or five mes-

sages a day, but by March, with the first Colossus still not fully operational, this had already risen to 20. More machines were ordered. Here too work was segregated: cryptographers (overwhelmingly male and upper middle class) directed female Colossus operators. Once chi wheel settings were known then psi and motor wheel settings were usually performed in the Testery. Using linguistic knowledge and craft skills, rather than mechanical aids, it took about an hour to identify the correct settings for a message if things went well. Bletchley Park had, after all, deliberately recruited champion crossword solvers. Colossus could handle this job, but usually helped out only when spare machine time was available. Once all settings were known women, mostly from working class backgrounds, operated machines to decrypt the messages. The Tunny messages were translated by a mixed-gender group of middle class university graduates, analyzed by male intelligence specialists, and filed by women in a massive repository indexed with punched cards.^b

Tutte realized that his statistical method could also, given a sufficiently long message, be used to break wheels. The technique was called "rectangling" because it involved making a large table tabulating every possible position of two wheels. The first two chi wheels repeated themselves every 1,271 characters, meaning a very long message might contain a dozen or more characters encoded with each possible wheel combination. Whenever the psi wheels did not rotate, the sum of the deltas from the encrypted message would be the same as the sum of the deltas on the corresponding channel of the chi wheel sequence and the plain text. Because the deltas of the first two channels of plain text were disproportionately likely to sum to zero, this exposed information about the wheel patterns. The encrypted message deltas tallied in each cell of the table gave clues about the sum of the deltas of the two corresponding code wheels. A high score

^b The gender and class background of the different jobs is taken from Christopher Smith, *The Hidden History of Bletchley Park*, Palgrave, 2015. The work of Colossus operators is discussed in Abbate, J. *Recoding Gender: Women's Changing Participation in Computing*. MIT Press, Cambridge, MA, 2012.

pointed to a delta total of one, a low score to a total of zero. Codebreakers worked from estimates of the delta totals back to the separate deltas on each wheel, using a technique similar to the solving of a “magic rectangle.” These candidate wheel deltas were then set up on Colossus, and varied over many runs to maximize fit with other messages using the same wheel patterns. Later versions of Colossus included an extra control panel, making it easier to set up and test candidate wheel patterns without going around to the back of the machine. Once the deltas for each wheel were established deriving the raw wheel patterns was trivial in comparison. Wheel breaking, like wheel setting, no longer relied on slips by the German operators. This was just as well, because in July 1944 they started changing all wheel patterns daily.

Colossus in Context

Turing is often misidentified by journalists as the creator of Colossus. As I discussed in a previous column, the “Matthew Effect” means that all the credit for any technical achievement tends to go to the most famous person tangentially associated with it. *The Imitation Game* further confuses this: “I want Christopher to be ... a digital computer” Cumberbatch says of his Bombe, after being asked whether he is trying to build a Turing machine.

This seems to deliberately blur the Bombe with Colossus, which has often been called “the first programmable electronic computer,” though Flowers himself was ambivalent about calling it a computer in his talks and articles. He saw it as a “processor,” part of the evolution of digital bit stream manipulation toward his dream of a fully electronic telephone exchange. After working with Mark Priestley on a detailed analysis of Colossus I agree with Flowers. Colossus could be set up to combine input bitstreams from tape and the simulated code wheels in many different ways, running the selected inputs through a network of logic gates. This gave it the flexibility to carry out many different Tunny-related attacks. Because it did not include any hardware for arithmetic calling it a computer confuses things, whether we go by today’s meaning of the word or 1940s definitions. Its con-

Newly declassified information gives us a better sense of just how impressive Colossus was.

figurable logic was stateless, used only to combine the current set of inputs to increment, or not increment, each of its five counters. The counters retained state from one tape character to the next, but the logic networks could not access this information. Neither was it programmable, as its basic sequence of operations was fixed (though many parameters could be set with switches). Colossus was a milestone in the development of digital electronics, incorporating many of the circuits and hardware features later used to build computers. But it had a quite different purpose, and different architecture, from an electronic computer.

Newly declassified information gives us a better sense of just how impressive Colossus was. During the war the U.S. commissioned dozens of different specialized code-breaking machines from prominent companies and research centers such as IBM, NCR, Bell Labs, and MIT. Almost without exception the most technologically ambitious projects, such as microfilm and electronic memories, were late, abandoned before completion, or proved so unreliable as to be useless. Even some of the more technologically conservative projects, such as a million-dollar giant Bombe built by Bell Labs, fizzled in use because they were constructed without a good grasp of code-breaking practice.^c Likewise, pretty much all post-war electronic computers, which like Colossus depended on making thousands of vacuum tubes work together, were plagued by reliability problems and took far longer to finish

than expected. The more I look at those other projects of the 1940s the more amazed I am that the Dollis Hill team had the first Colossus installed and working at Bletchely Park less than a year after pitching it to Newman.

Remembering and Forgetting

Back to *The Imitation Game*, which never mentions Tunny, Flowers, Colossus, or Tutte. The end of the movie takes place several years after the war. Cumberbatch’s Turing is coming apart mentally after being sentenced to compulsory hormone treatments intended to contain his homosexual urges. He bumbles around his house in a bathrobe, drifting toward suicide, distracting himself only by building another Christopher. To send the audience out on a more cheerful note, Joan Clarke, his friend and former fiancée, visits to point out that she had “just taken a train through a city that would not exist if it wasn’t for you. I bought a ticket from a man who would likely be dead if it wasn’t for you.”

When victory was declared, Cumberbatch boasted that “the war was really just a half-dozen crossword enthusiasts in a tiny village in the south of England. Was I God? No. Because God didn’t win the war.” I’m sure that the real Turing never dismissed the millions who fought and died in this way. He wasn’t an idiot. But beneath the film’s sepia nostalgia lies a very modern delusion: established institutions (such as government and the military) are so dysfunctional, and ordinary people so irrelevant, that tech geniuses must secretly seize power. To put that another way, Turing wasn’t Peter Thiel.

Tutte made the same kind of dazzling intellectual contributions as Turing to the success of Bletchley Park. But brilliance alone does not win one the kind of fame that culminates in a glossy scientific biopic—an extremely rare honor. After earning his Ph.D. in 1948 Tutte followed the trajectory of his wartime work into a distinguished mathematical career focused on graphs and matroids. He is well remembered in those fields, and his wartime contribution is documented, for those who care, in several histories of Bletchley Park. Tragedy and drama attracted Hollywood to the lives of Alan Turing, Stephen Hawking (*The Theory*

^c These projects are described in Burke, C.B. *It Wasn’t All Magic: The Early Struggle to Automate Cryptanalysis, 1930s–1960s*. National Security Agency, 2002.

of Everything), and John Nash (*A Beautiful Mind*). Contentment is not as exciting to watch as tragedy, but most of us would prefer it for ourselves. I don't think Tutte minded this relative neglect—he appears to have lived a long, happy, and active life full of scholarly accomplishments and recognition, including appointment as a Fellow of the Royal Society and even the posthumous naming of an asteroid.

Flowers was less satisfied than Tutte with his career, never entirely realizing his vision for a fully electronic telephone exchange. His relative obscurity reflects a larger tendency to remember conceptual breakthroughs and ideas, which fit the stereotype of individual genius, rather than system building accomplishments, which are inherently team based. He lived to experience a small measure of fame for his work on Colossus, but seemed bitter, complaining that his career was over before he could benefit from this newfound respect. His reputation has continued to grow, and in the past few years a postage stamp and a display in the Science Museum in London

have celebrated Colossus as a founding contribution to digital technology.

The real story of Bletchley Park is one of teamwork on a massive scale, in which good management and effective procurement were just as important as individual scientific genius. Tutte and Flowers did not win the war by themselves, any more than Turing did. But, whether or not the Colossus story ever makes it to the big screen, each played an outsize role in the defeat of fascism. **C**

Further reading

Copeland, Jack.

Colossus: The First Electronic Computer. Oxford University Press, New York, 2006. A compendium of analysis, reprinted historical documents and memoir. Contains Tutte's own description of his breakthroughs against Tunny.

Gannon, Paul.

Colossus: Bletchley Park's Greatest Secret (Atlantic Books, 2006). The clearest overall history of Colossus.

Hodges, Andrew.

Alan Turing: The Enigma. Princeton

University Press, Princeton, NJ, 2014. *The Imitation Game* was nominally based on Hodges' landmark biography of Turing, but don't let that put you off. He's maintained a tactful silence about the movie.

McKay, Sinclair.

The Secret Lives of Codebreakers: The Men and Women Who Cracked the Enigma Code at Bletchley Park. Gives a fascinating window onto the social history of Bletchley Park and the experiences of the people who worked there.

Randell, Brian.

"The Colossus." In Metropolis, N., Howlett, J. and Rota, G.-C. eds., Academic Press, New York, 1980, 47-92. This paper broke the public silence surrounding Colossus, at a time when the British government was still actively suppressing the release of information.

Welchman, Gordon.

The Hut Six Story: Breaking the Enigma Codes (McGraw Hill, 1982). The most substantial first person account of Bletchley Park, focused on Enigma rather than Colossus.

Thomas Haigh (thomas.haigh@gmail.com) is a Visiting Professor at Siegen University and an Associate Professor at the University of Wisconsin—Milwaukee. He is the primary author of *ENIAC In Action* (MIT Press, 2016). Read more at <http://www.tomandmaria.com/tom>

Copyright held by author.

FACULTY POSITIONS IN COMPUTER SCIENCE

(Full, Associate and Assistant Professor)

- Data Mining and analytics with emphasis on Big Data
- Machine Learning with emphasis on Deep Learning
- Artificial intelligence
- Computer Systems and Emerging Architectures (GPU's, FPGA's, etc)
- High performance computing - Computer security

Committees: 1. For Data Mining, Machine Learning AI, Emerging Architectures and Security: Panos Kalnis, Mootaz Elnozahy, Peter Wonka, Wolfgang Heidrich, Xin Gao, Marco Canini, Khaled Salama 2. For High Performance Computing: Committee appointed by David Keyes

Applications will be considered until the positions are filled but not later than April 15, 2017



CEMSE Computer, Electrical and Mathematical Sciences and Engineering



Please apply via the **cemse.kaust.edu.sa** employment site

KAUST offers superb research facilities, including the 5 Petaflop/s Shaheen-2 supercomputer



KAUST generous assured research funding and internationally competitive salaries

Viewpoint

Artificial Intelligence: Think Again

Social and cultural conventions are an often-neglected aspect of intelligent-machine development.

THE DOMINANT PUBLIC narrative about artificial intelligence is that we are building increasingly intelligent machines that will ultimately surpass human capabilities, steal our jobs, possibly even escape human control and kill us all. This misguided perception, not widely shared by AI researchers, runs a significant risk of delaying or derailing practical applications and influencing public policy in counterproductive ways. A more appropriate framing—better supported by historical progress and current developments—is that AI is simply a natural continuation of longstanding efforts to automate tasks, dating back at least to the start of the industrial revolution. Stripping the field of its gee-whiz apocalyptic gloss makes it easier to evaluate the likely benefits and pitfalls of this important technology, not to mention dampen the self-destructive cycles of hype and disappointment that have plagued the field since its inception.

At the core of this problem is the tendency for respected public figures outside the field, and even a few within the field, to tolerate or sanction overblown press reports that herald each advance as startling and unexpected leaps toward general human-level intelligence (or beyond), fanning fears that “the robots” are coming to take over the world. Headlines often tout noteworthy engineering accomplishments in a context suggesting they constitute unwelcome assaults on human



uniqueness and supremacy. If computers can trade stocks and drive cars, will they soon outperform our best sales people, replace court judges, win Oscars and Grammys, buy up and develop prime parcels of real estate for their own purposes? And what will “they” think of “us”?

The plain fact is there is no “they.” This is an anthropomorphic conceit borne of endless Hollywood blockbusters, reinforced by the gratuitous inclusion of human-like features in public AI technology demonstrations, such

as natural-sounding voices, facial expressions, and simulated displays of human emotions. Each of these techniques has valuable application to human-computer interfaces, but not when their primary effect is to fool or mislead. Attempts to dress up significant AI accomplishments with humanoid flourishes does the field a disservice by raising inappropriate questions and implying there is more there than meets the eye. Was IBM’s Watson pleased with its “Jeopardy!” win? It sure looked like it. This made for great

television, but it also encouraged the audience to overinterpret the actual significance of this important achievement. Machines don't have minds, and there is precious little evidence to suggest they ever will.

The recent wave of public successes, remarkable as they are, arise from the application of a growing collection of tools and techniques that allow us to take better advantage of advances in computing power, storage, and the wide availability of large datasets. This is certainly great computer science, but it is not evidence of progress toward a superintelligence that can outperform humans at any task it may choose to undertake. While some of the new tools—most notably in the field of machine learning—can be broadly applied to classes of tasks that may appear unrelated to the non-technical eye, in practice they often rely upon certain common attributes of the problem domains, such as enormous collections of examples in digital form. High-speed trading algorithms, tracking objects in videos, and predicting the spread of infectious diseases all rely on techniques for finding subtle patterns in noisy streams of real-time data, and so many of the tools applied to these apparently diverse tasks are similar.

We are certainly using machines to perform all sorts of real-world tasks that people perform using their native intelligence, but this does not mean the computers are intelligent. It merely means there are other ways to solve these problems. People and computers can play chess, but it is far from clear that they do it the same way. Recent advances in machine translation are remarkably successful, but they rely more on statistical correlations gleaned from large bodies of concorded texts than on fundamental advances in the understanding of natural language.

Machines have always automated tasks that previously required human effort and attention—both physical and mental—usually by employing very different techniques. And they often do these tasks better than people can, at lower cost, or both—otherwise they would not be useful. Factory automation has replaced myriad highly skilled and highly trained workers, from sheet metal workers to coffee tasters. Arithmetic

Machines don't have minds, and there is precious little evidence to suggest they ever will.

problems that used to be the exclusive domain of human “calculators” are now performed by tools so inexpensive they are given away as promotional trinkets at trade shows. It used to take an army of artists to animate Cinderella's hair, but now CGI techniques render Rapunzel's flowing locks. These advances do not demean or challenge human capabilities; instead they liberate us to perform ever more ambitious tasks.

Some pundits warn that computers in general, and AI in particular, will lead to widespread unemployment. What will we do for a living when machines can perform nearly all of today's jobs? A historical perspective reveals a potential flaw in this concern. The labor market constantly evolves in response to automation. Two hundred years ago, more than 90% of the U.S. labor force worked on farms. Now, barely 2% produce far more food at a fraction of the cost. Yet, everyone isn't out of work. In fact, more people are employed today than ever before, and most would agree their jobs are far less taxing and more rewarding than the backbreaking toil of their ancestors. This is because the benefits of automation make society wealthier, which in turn generates demand for all sorts of new products and services, ultimately expanding the need for workers. Our technology continually obsoletes professions, but our economy eventually replaces them with new and different ones. It is certainly true that recent advances in AI are likely to enable the automation of many or most of today's jobs, but there is no reason to believe the historical pattern of job creation will cease.

That's the good news. The bad news is that technology-driven labor mar-

ket transitions can take considerable time, causing serious hardships for displaced workers. And if AI accelerates the pace of automation, as many predict, this rapid transition may cause significant social disruption.

But which jobs are most at risk? To answer this question, it's useful to observe that we don't actually automate jobs, we automate tasks. So whether a worker will be replaced or made more productive depends on the nature of the tasks they perform. If their job involves repetitive or well-defined procedures and a clear-cut goal, then indeed their continued employment is at risk. But if it involves a variety of activities, solving novel challenges in chaotic or changing environments, or the authentic expression of human emotions, they are at far lower risk.

So what are the jobs of the future? While many people tend to think of jobs as transactional, there are plenty of professions that rely instead on building trust or rapport with other people. If your goal is to withdraw some spare cash for the weekend, an ATM is as effective as a teller. But if you want to secure an investor to help you build your new business, you won't be pitching a machine anytime soon.

This is not to say that machines will never sense or express emotions; indeed, work on affective computing is proceeding rapidly. The question is how these capabilities will be perceived by users. If they are understood simply as aids to communication, they are likely to be broadly accepted. But if they are seen as attempts to fake sympathy or allay legitimate concerns, they are likely to foster mistrust and rejection—as anyone can attest who has waited on hold listening to a recorded loop proclaim how important their call is. No one wants a robotic priest to take their confession, or a mechanical undertaker to console them on the loss of a loved one.

Then there are the jobs that involve demonstrations of skill or convey the comforting feeling that someone is paying attention to your needs. Except as a novelty, who wants to watch a self-driving racecar, or have a mechanical bartender ask about your day while it tops up your drink? Lots of professions require these more social skills, and the demand for them is only going

Model Learning**Smart Machines
Are Not A Threat
to Humanity****AI Dangers:
Imagined and Real****Computing History
Outside the U.K. And U.S.:
Some Selected
Landmarks from
Continental Europe****Copyright Enforcement
in the Digital Age****A Messy State
of the Union:
Taming the Composite
State of Machines of TLS****Life Beyond Distributed
Transactions****BBR:
Congestion-Based
Congestion Control****Are You
Load Balancing
Wrong?**

Plus the latest news about closing the back door, predictive policing, and secure quantum communications.

So the robots are certainly coming, but not in the way most people think.

to grow as our disposable income increases. There's no reason in principle we can't become a society of well-paid professional artisans, designers, personal shoppers, performers, caregivers, online gamers, concierges, curators, and advisors of every sort. And just as many of today's jobs did not exist even a few decades ago, it is likely a new crop of professions will arise that we can't quite envision today.

So the robots are certainly coming, but not quite in the way most people think. Concerns that they are going to obsolete us, rise up, and take over, are misguided at best. Worrying about superintelligent machines distracts us from the very real obstacles we will face as increasingly capable machines become more intricately intertwined with our lives and begin to share our physical and public spaces. The difficult challenge is to ensure these machines respect our often-unstated social conventions. Should a robot be permitted to stand in line for you, put money in your parking meter to extend your time, use a crowded sidewalk to make deliveries, commit you to a purchase, enter into a contract, vote on your behalf, or take up a seat on a bus? Philosophers focus on the more obvious and serious ethical concerns—such as whether your autonomous vehicle should risk your life to save two pedestrians—but the practical questions are much broader. Most AI researchers naturally focus on solving some immediate problem, but in the coming decades a significant impediment to widespread acceptance of their work will likely be how well their systems abide by our social and cultural customs.

Science fiction is rife with stories of robots run amok, but seen from an engineering perspective, these are de-

sign problems, not the unpredictable consequences of tinkering with some presumed natural universal order. Good products, including increasingly autonomous machines and applications, don't go haywire unless we design them poorly. If the HAL 9000 kills its crewmates to avoid being deactivated, it is because its designers failed to prioritize its goals properly.

To address these challenges, we need to develop engineering standards for increasingly autonomous systems, perhaps by borrowing concepts from other potentially hazardous fields such as civil engineering. For instance, such systems could incorporate a model of their intended theater of operation, (known as a Standard Operating Environment, or SOE), and enter a well-defined "safe mode" when they drift out of bounds. We need to study how people naturally moderate their own goal-seeking behavior to accommodate the interests and rights of others. Systems should pass certification exams before deployment, the behavioral equivalent of automotive crash tests. Finally, we need a programmatic notion of basic ethics to guide actions in unanticipated circumstances. This is not to say machines have to *be* moral, simply that they have to *behave* morally in relevant situations. How do we prioritize human life, animal life, private property, self-preservation? When is it acceptable to break the law?

None of this matters when computers operate in limited, well-defined domains, but if we want AI systems to be broadly trusted and utilized, we should undertake a careful reassessment of the purpose, goals, and potential of the field, as least as it is perceived by the general public. The plain fact is that AI has a public relations problem that may work against its own interests. We need to tamp down the hyperbolic rhetoric favored by the popular press, avoid fanning the flames of public hysteria, and focus on the challenge of building civilized machines for a human world. **C**

Jerry Kaplan (jerrykaplan@stanford.edu) is a visiting lecturer in computer science at Stanford University and a fellow at the Stanford Center for Legal Informatics. His latest book is *Artificial Intelligence: What Everyone Needs to Know* (Oxford University Press, 2016).

Copyright held by author.

Viewpoint

Effects of International Trafficking in Arms Regulations Changes

Considering the impact of recent ITAR changes to the U.S. software industry and software education.

WHILE IT IS an old adage that the pen is mightier than the sword, the U.S. Department of State may have taken this concept a bit far too in classifying your email client as a munition. On June 3, 2015, the U.S. Department of State released new proposed International Trafficking in Arms Regulations (ITAR) rules.⁸ Unlike many of the previous rulemaking releases that were part of this process, this release dealt with general terms that underlie the regulations in many other sections. As part of these, seemingly innocuous, changes, they replaced several definitions. This Viewpoint considers the impact of these changes on the U.S. software industry and software education.

The U.S. has a very real need to protect certain information. Some of this information is so sensitive that it can be made available to only select individuals within or working for the government. A classification mechanism exists for this government-originated information that restricts it to only authorized users. The location, configuration, and capabilities of military assets and other similar information fall under this regime. Other information and certain goods are also seen as providing the U.S. strategic advantage and thus are restricted to only U.S. and authorized foreign



An attendee of the U.S. Military Academy using the West Point Simulation Center for familiarization with military weapons and tactics.

use. Two regimes exist for controlling this information and these goods: the ITAR⁶ and the Export Administration Regulations (EAR).² The former covers items that have a definitive military use (or for which military use is likely); the later covers less harmful goods. ITAR items are regulated by the U.S. Department of State, while EAR items are regulated by the U.S. Department of Commerce. The two regimes have many similarities; however, the process, certain exemptions and the penalties for violation can differ significantly.

The ITAR changes have drawn no shortage of criticism⁷ with some arguing, for example, that the regulations have made the U.S. uncompetitive in the space and aerospace industries. Others³ have asserted that they may violate freedom of speech protections embodied in the U.S. Constitution. Previous work⁵ has highlighted the impact that these regulations may have on small businesses, academic institutions, and individuals and proposed a safe harbor be created for these groups (with the space and aerospace industries specifically considered).

This Viewpoint bypasses the often discussed question of what the role of export control should be in our society and whether current regulations meet this goal and instead focuses on the impact of several changes to the ITAR that have been recently proposed. It is important to note that these changes are simply being proposed at this point, precisely so that a dialog (like the one hopefully started here) can be conducted to inform the agency's final actions.

The proposed changes⁸ were released as a proposed rule. The formal public comment period on this closed on August 3, 2015. The agency could, conceivably, still consider informal public comments. The next step in this process would, typically, be the release of a final rule (or the agency may opt to solicit feedback again before making a final rule, depending on a variety of factors). Of course, further changes to these rules could be initiated by the agency (at the request of the public or at its own initiative) or be necessitated by congressional action.

What Changed

In the June 3, 2015 proposed rulemaking,⁸ several critical aspects of ITAR were changed:

- Software was moved from falling under technical data to being a defense article. While the associated commentary states that this is “not a substantive change,” this may be incorrect as certain sections of ITAR (for example, defense service regulation) apply differently to technical data versus defense articles.⁹

- The definition of public domain was changed significantly. Specifically “technical data or software, whether or not developed with government funding, is not in the public domain if it has been made available to the public without authorization from” one of several enumerated federal agencies. This effectively curtails the ability of private-sector research firms to use the public domain exemption as a pathway to ITAR exemption.¹⁰

- Private entities conducting federally funded research can qualify under the fundamental research exemption.¹¹

Analysis of the Impact and Importance of these Changes

The proposed changes will be analyzed

from three perspectives: the U.S. software industry (with a particular focus on small developers), computer science educators, and the scientific enterprise in general.

U.S. software industry. From the perspective of the U.S. software industry, the changes are potentially problematic. Several sections are vague and could pull large segments of commercial software under the control of ITAR (meaning that sales, even in the U.S., would have to be closely monitored to ensure they were only to U.S. persons or a license would need to be obtained). For example, Category IX includes:¹²

Software and associated databases not elsewhere enumerated in this subchapter that can be used to model or simulate the following:

- Trainers enumerated in paragraph (a) of this category;
- Battle management;
- Military test scenarios/models

This would seem to encompass many battle-type games (notably, these types of games have been demonstrated to be usable for troop training⁴); however, the broad language “software ... that can be used” could ensnare operating systems and other benign applications that might have a supporting role in such modeling or simulation. Presumably, it is not the intent of the State Department to regulate the entire domestic software industry. However, fear and uncertainty could have a chilling effect on domestic development, sales, and exports. This may be particularly problematic for small firms that cannot afford to hire attorneys to advise them on ITAR compliance. Certain types of cyberphysical systems (for example, “vehicle management computers”¹³ and security systems¹⁴) and their associated software may be particularly difficult to find an exemption for.

Computer science educators. Fortunately, a note¹⁵ carves out an exemption for many educational activities, exempting “instruction in general scientific, mathematical, or engineering principles commonly taught in schools, colleges, and universities.” However, what may be implicated under the new regulations is the assistance that could otherwise be provided to students working on their own efforts (for example, in the context of project-based learning.¹⁵ The new language may also

The ITAR changes have drawn no shortage of criticism, with some arguing that the regulations have made the U.S. less competitive.

impair faculty aid to student start-ups, spin-outs and industry collaboration. Because software is now defined as a defense article, much of this aid may qualify as a defense service.

For example, while a university faculty member might be able to provide foreign national students instruction on a topic, he or she could be precluded (absent application and approval) from providing feedback on an extracurricular project. Depending on how “instruction” is defined, this prohibition could even be taken to extend to providing feedback and guidance on class-related project based learning activities. If these same foreign national students sought to start a small business related to the material they had learned, the faculty member would not be able to assist them without securing approval.

The research enterprise. While research in academia enjoys similar protection to before (embodied in a new section 120.49), industry loses a key exemption that has allowed them to participate in scientific discourse: the public domain exemption. While the university exemption is now expanded to cover federally funded research in industry, materials meeting the previous definition of being in the public domain (note that public domain is defined differently in ITAR than under copyright law) do not qualify for an exemption unless authorized. Authorizing agencies include the Directorate of Defense Trade Controls, The Department of Defense's Office of Security Review, “relevant U.S. government contracting” entities “with authority” or “another U.S. government official with authority.”¹⁴ Making technical data

publicly available (a key component of the public domain exemption) over the Internet is now defined as a type of export (irrespective of whether it reaches a foreign person), which may impair growth in Internet software delivery and cloud services.¹⁷

This change could prospectively reduce the participation of industry in conferences and other (for example, journal) publications. The need for agency approval may increase the expense incurred by members of the public and small and large businesses for participating in presentations and publications. This change, thus, may reduce privately funded research outside of the university environment. A small business that has developed an innovative technology, for example, may decide that the risk created by non-review and the cost and delay of government review make publication of their work untenable. This firm may, thus, decide not to allow publication, impairing the use of the discovery by the broader scientific community and the career advancement of the scientist who made the discovery.

Pathway to Correction

A logical question, following from the foregoing identified problems, is how to prospectively correct these (possibly unintended) impacts of the ITAR changes. This would seem to have several prospective approaches.

The first is to create a separate classification for software that is neither technical data nor a defense article. This would allow regulations to be created and applied specifically to software (which is arguably different from both other categories and doesn't fit perfectly in either) after appropriate consideration of the needs related specifically to defense versus general-use software.

The second issue is the change to the public domain definition. Setting up the government as the clearinghouse of what is and is not in the public domain (particularly without any consideration to items previously generally available) is inherently problematic. The old approach (which allowed a pathway for almost anything that wasn't government funded or classified to enter the public domain) was probably overly broad; however, the proposed approach is problematically burdensome and restrictive. In the short term, changing

the proposed rule to explicitly include anything previously meeting the public domain definition under the current language (before the change was made) in the public domain definition and anything substantially similar (in that it proposes no new export control concerns) to items currently in the public domain would eliminate some problems. More broadly, the notion of what information and software should and should not be controlled by the Department of State should be a topic for greater debate, as there is clearly a balance that needs to be struck.

Finally, enacting a safe harbor (as proposed by Straub and Vacek⁵) could be an important step to keeping small businesses and individuals innovating. The safe harbor proposed suggested that in most cases the government must show actual harm (as opposed to just a, prospectively technical, rules violation) to be successful in a prosecution or civil claim.

Conclusion

This most recent set of proposed changes to the ITAR attempt to clarify several topics. They better define the concept of what is and is not protected by the exemption related to fundamental research (likely narrowing the definition somewhat from what it may be perceived as, at present). They add additional sections codifying definitions for development, production, release and retransfer. Further, they add and expand several types of export definitions.

Problematically, however, they also regulate software development activities that are not defense directed. They raise consideration of whether certain instructional activities may constitute defense services, with regard to software (now a defense article) development and they prospectively impair the dissemination of research by corporate researchers (and, prospectively, harm collaboration between corporate scientists and their university counterparts).

The foregoing illustrates the necessity for the computer science and the greater scientific community to get more involved with policy development. In the short term, it is important to clarify the State Department's intent with regards to the foregoing (much of which may be unanticipated or unintended consequences). Continued

proactive involvement will also facilitate the shaping of this important topic into the future. C

References

1. Blumenfeld, P.C. et al. Motivating project-based learning: Sustaining the doing, supporting the learning. *Educational Psychologist* 26 (1991), 369–398.
2. Export Administration Regulations, Code of Federal Regulations: 15 CFR 730–774.
3. Gold, M. Thomas Jefferson, we have a problem: The unconstitutionality nature of the U.S.'s aerospace export control regime as supposed by *Bernstein v. U.S. Department of Justice*. *Cleveland State Law Review* 57 (2009).
4. Proctor, M. Are officers more reticent of games for serious training than enlisted soldiers? *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* 5 (2008), 179–196 (2008).
5. Straub, J. and Vacek, J. Reforming regulation of basic and small business research and education in Space technologies under ITAR (International Traffic in Arms Regulations) and EAR (Export Administration Regulations). *Space Law Journal* 39, 2 (2014).
6. The International Traffic in Arms Regulations, Code of Federal Regulations: 22 CFR 120–130.
7. Tushe, N. U.S. export controls: Do they undermine the competitiveness of U.S. companies in the transatlantic defense market. *Public Contract Law Journal* 41 (2011–2012).
8. U.S. Department of State. International traffic in arms: Revisions to definitions of defense services, technical data, and public domain; definition of product of fundamental research; electronic transmission and storage of technical data; and related definitions. *Federal Register* 80, 106, (June 3, 2015), 31525–31538.
9. U.S. Department of State. International traffic in arms: Revisions to definitions of defense services, technical data, and public domain; definition of product of fundamental research; electronic transmission and storage of technical data; and related definitions. *Federal Register* 80, 106 (June 3, 2015), 31526.
10. U.S. Department of State. International traffic in arms: Revisions to definitions of defense services, technical data, and public domain; definition of product of fundamental research; electronic transmission and storage of technical data; and related definitions, proposed section 120.11(b). *Federal Register* 80, 106 (June 3, 2015), 31535.
11. U.S. Department of State. International traffic in arms: Revisions to definitions of defense services, technical data, and public domain; definition of product of fundamental research; electronic transmission and storage of technical data; and related definitions, proposed section 120.49(a)(2). *Federal Register* 80, 106 (June 3, 2015), 31536.
12. U.S. Munitions List. Code of Federal Regulations: 22 CFR 121.1, Category IX(b)(4).
13. U.S. Munitions List. Code of Federal Regulations: 22 CFR 121.1, Category VII(g)(12).
14. U.S. Munitions List. Code of Federal Regulations: 22 CFR 121.1, Category XIII(b).
15. U.S. Department of State. International traffic in arms: Revisions to definitions of defense services, technical data, and public domain; definition of product of fundamental research; electronic transmission and storage of technical data; and related definitions, proposed section 120.9(note to paragraph(a))(9). *Federal Register* 80, 106 (June 3, 2015), 31534.
16. U.S. Department of State. International traffic in arms: Revisions to definitions of defense services, technical data, and public domain; definition of product of fundamental research; electronic transmission and storage of technical data; and related definitions, proposed section 120.11(b)(1–4). *Federal Register* 80, 106 (June 3, 2015), 31535.
17. U.S. Department of State. International traffic in arms: Revisions to definitions of defense services, technical data, and public domain; definition of product of fundamental research; electronic transmission and storage of technical data; and related definitions, proposed section 120.17(a)(7). *Federal Register* 80, 106 (June 3, 2015), 31535.

Jeremy Straub (jeremy.straub@ndsu.edu) is an assistant professor in the Department of Computer Science at North Dakota State University.

Copyright held by author.

Article development led by **acmqueue**
queue.acm.org

Don't "win." Resolve.

BY KATE MATSUDAIRA

Resolving Conflict

RECENTLY, I WAS talking to one of my employees about one of my least/most favorite topics in the whole world: conflict.

I am conflicted about conflict. On one hand, I hate it. Hearing people disagree, even about minor things, makes me want to run through the nearest wall and curl up under my bed until it's over.

On the other hand, when it happens, I always want to get into it.

I think that urge to jump in and get involved actually comes from my discomfort with conflict; I hate it so much that when it comes up, I just want to dive in so it can be over as soon as possible.

I have this need to help everyone understand each other's point of view, show them what they have in common, and make it so the conflict is just over.

By leaning into conflict, rather than trying to avoid it, I think you can often actually get it over with faster. And it's a pretty good thing to be known as a person who can help everyone get on the same page and get back to being productive.

How do you feel about conflict? Especially conflict at work?

In a perfect world, we would all get along with our coworkers and bosses all the time. Unfortunately, we don't live in a perfect world.

While most of us make our best efforts to avoid conflict at work, occasionally it is unavoidable. Here are some of my best tips on how to make all of your conflicts in the workplace healthy and (hopefully) productive, so you can move on and get back to what really matters.

1. Give Up on the Idea of "Winning"

The best way to win an argument is to let go of the idea that you actually have something to "win."

Winning, in this case, doesn't mean getting your way or showing the opposition how they are wrong. Instead, it means being the person who helps everyone get on the same page so everyone can move forward.

With most technical decisions there aren't "right answers;" there are only different approaches that all have pros and cons. Getting aligned in the decision-making process and determining which trade-offs are acceptable is better than trying to demonstrate who is right.

If you want to be successful (whether as a leader or a senior engineer), you need to be someone who can look at the big picture, assess how to move forward, and then get everyone working on the same page again. That is true leadership and what most managers value in great employees.

2. Start by Looking for Common Ground

At the heart of many workplace conflicts is often a common goal. Two people disagreeing over strategy might have the shared goal of wanting to execute a project at the highest quality possible. Their conflict isn't as deep as it might look from the outside; really, they already agree on the important parts and are just fighting about details.



For example, this happens a lot when you have operations and software development teams with competing priorities: the ops team wants to minimize change and risk (and thereby operations), and the dev team wants to ship their features as fast as possible. The reality is both teams should be focused on the best thing for the business and the customer—which is likely somewhere in the middle.

When you can see what you have in common with the other side, then you can start to sort out the facts and determine the key priorities.

- Why does each person think what they think?

- Is there outside information that could influence or persuade them otherwise?

- Why do you think what you think?

Finding common ground makes

compromise easier, since the other person's perspective can feel relatable and reasonable, and you realize that person is more like you than you thought.

3. Don't Blow Up (And If You Do, Leave)

It's really difficult to agree with or give in to someone you're mad at. The more worked up you are, the more defensive you get and the less listening you do.

Not only will you be damaging the relationship if you blow up at the other person, but you also won't be getting any closer to a resolution. It is a waste of everyone's time. It is easy to lose your temper when you get frustrated or feel like you aren't being heard (or understood)—but this is such an important lesson to remember. You have to keep your cool.

If either you or the other person is losing it, walk away. Explain that you need to take a break, and come back later. If need be, apologize and then come back to the question when you have a cooler head and are more likely to be thinking logically.

4. Focus On the Facts (Not the Perceptions)

You might think you know why a person has a certain opinion or why they do their work a certain way, but don't assume. Chances are, you are wrong, and, besides, nobody takes kindly to hearing what other people think of them (especially if they are worked up and frustrated).

The best thing you can do during a conflict is to focus on the facts. Speak only for yourself. Avoid saying things like, "We all think ..." or "You're just

saying that because ... ” Instead, talk about your experience, your knowledge, and the facts at hand.

Try to take as much emotion and projection out of it as possible, and just look at what is in front of you.

- ▶ What is the goal?
- ▶ What are the possible solutions?
- ▶ How can we measure each of them?
- ▶ Do we have any experiences or resources we can draw on to get more information?
- ▶ How can we reach a compromise that acknowledges everyone's needs?

5. Repeat the Other Person's Words Back to Them

In an argument, it can be tempting to reiterate your position again and again. We all want to feel that we are being heard, and when emotions are hot, it is difficult to think beyond our own opinions.

The more you can listen to the other person, however, the more you will make them feel heard (which lowers the level of conflict) and the more you will understand their perspective (which will help you uncover what you need to know to find a workable resolution).

Repeating the other person's words back to them is a great way to do this.

When someone finishes making a point, you can acknowledge that you heard it by saying, “OK, that makes sense. Just to make sure I completely understand, you are saying ... ” and repeat their key points.

When people feel they are being heard, they are more likely to compromise because they feel like their perspective is being taken into account in the decision. (And feeling like they weren't being heard is likely what started the conflict in the first place.)

6. Stick With It and Seek a Conclusion

It can be tempting to implement the silent treatment or simply walk away when someone disagrees with you, but it's important to see the conflict through to resolution.

As painful as that might sound, imagine the alternative: seething frustration that drags on for hours, days, or even years, and damages your relationships with coworkers—and maybe even your reputation, if

When people feel they are being heard, they are more likely to compromise because they feel like their perspective is being taken into account in the decision.

the blowup was big enough or causes enough long-term damage.

One bad interaction can turn into a bad relationship, which can have wide-reaching negative impacts on your career. Better to get the situation resolved now, so you can all move on.

It might be tempting to throw up your hands and say, “We'll never agree!” but it is better to seek a conclusion to the conflict than just to accept it. Maybe you ultimately will decide you and this other person just have to “agree to disagree,” but it is better to have that be a mutual decision than for one of you to walk away.

Remember, It's About Resolution, Not Winning

You don't always have to agree 100% in order to perform well in your job. Don't focus so much on winning that you turn into a sore loser if someone else appears to come out on top in the conflict.

Even if you don't get your way, remember that at the end of the day, it is your job to be aligned with your team and do great work.

If you pout and phone in your work because you didn't get your way, people will notice and they will remember—and that will make it even more difficult to get your way in the future.

If, however, you can work through a conflict, be a great teammate, and still produce great work, then you will become a respected authority on your team. The longer your track record of successfully managing and negotiating conflict, the better it will serve you in the long run than winning one fight. 

Related articles on queue.acm.org

Sink or Swim, Know When It's Time to Bail
Gordon Bell

<http://queue.acm.org/detail.cfm?id=966806>

System Administration Soft Skills
Christina Lear

<http://queue.acm.org/detail.cfm?id=1922541>

Delegation as Art

Kate Matsudaira

<http://queue.acm.org/detail.cfm?id=2926696>

Kate Matsudaira (katemats.com) is the founder of her own company, Popforms. Previously she worked in engineering leadership roles at companies like Decide (acquired by eBay), Moz, Microsoft, and Amazon.

Copyright held by author.
Publications rights licensed to ACM. \$15.00



**Using OpenFlow and DevOps
for rapid development.**

BY JOSH BAILEY AND STEPHEN STUART

Faucet: Deploying SDN in the Enterprise

THE 2008 PUBLICATION of “OpenFlow: Enabling Innovation in Campus Networks” introduced the idea that networks (originally campus and enterprise networks) can be treated more like flexible software rather than inflexible infrastructure, allowing new network services and bug fixes to be rapidly and safely deployed.⁷

Since then many have shared their experiences using software-defined networking (SDN) and OpenFlow in wide area and data center networks, including at Google.¹⁰ This article returns to enterprise and campus networks, presenting an open source SDN controller for such networks: Faucet. The Faucet controller provides a “drop-in” replacement for one of the most basic network elements—a switch—and was created to easily bring the benefits of SDN to today’s typical enterprise network.⁵

SDN enables such safe and rapid development and deployment of network features through automated

testing of both hardware and software, without time-consuming manual lab testing. As described here, a complete control-plane upgrade can be done, while the network is running, in a fraction of a second.

Security of networks is a concern for all network operators and users. A zero-day attack on the network itself is especially worrisome because it can impact the security of all users and services on the network. Therefore, it is critical that network operators have a way of responding rapidly, both to deploy new security features or mitigate vulnerabilities in advance of an attack

Figure 1. Non-SDN and Faucet SDN comparison.

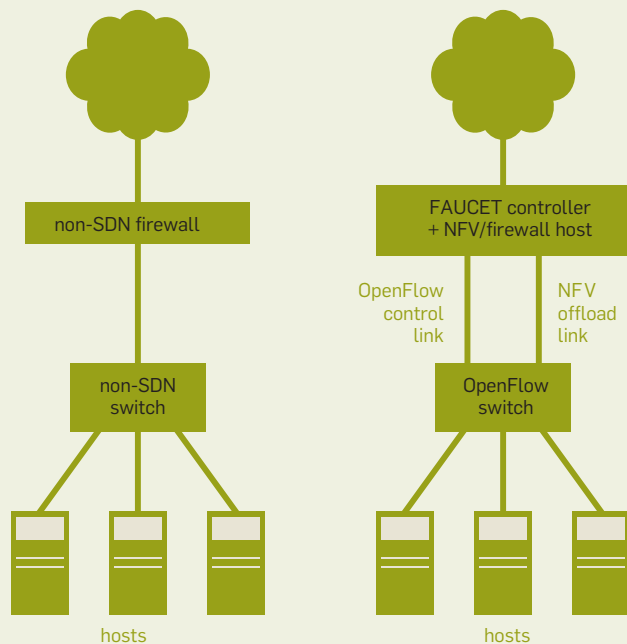


Figure 2. Faucet's all OpenFlow pipeline.

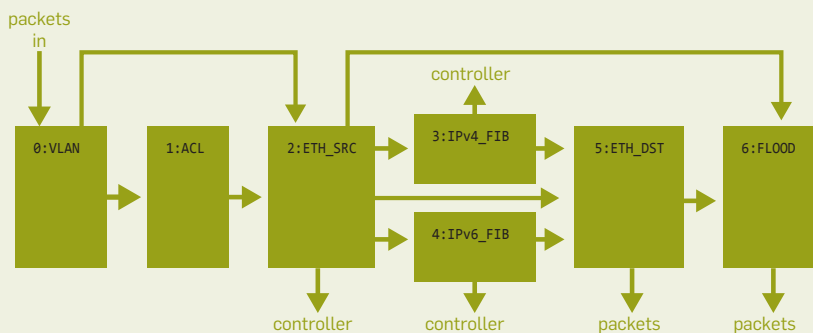


Figure 3. Using Faucet and Grafana to construct dashboards and run queries.



and to restore a network currently undergoing an attack as quickly and completely as possible, with as little risk as possible. SDN builds this ability to change and respond into the network itself at a very low level, which is beyond the reach of an external security device such as a firewall.

Faucet has been tested, and it performs. It has been deployed in various settings, including the Open Networking Foundation, which runs an instance of Faucet as its office network. Faucet delivers high forwarding performance using switch hardware, while enabling operators to add features to their networks and deploy them quickly, in many cases without needing to change (or even reboot) hardware. Furthermore, it interoperates with neighboring nonSDN network devices.

Faucet was built on the OpenFlow 1.3 standard.⁸ Without the availability of commercial hardware supporting this standard, it would not have been possible. Multiple vendors now ship hardware that supports OpenFlow 1.3, specifically with support for multiple flow tables and IPv6. To minimize vendor-specific logic in the controller, vendors were encouraged to support key features in the OpenFlow 1.3 standard in a consistent way. This reduced initial development and support cost, and it simplified bug reporting and automated testing.

While SDN as a technology continues to evolve and become even more programmable (for example, with the P4 programming language), Faucet and OpenFlow 1.3 hardware together are sufficient to realize benefits today. This article describes specifically how to take advantage of DevOps practices (“push on green”⁶) to develop and deploy features rapidly. It also describes several practical deployment scenarios, including firewalling and network function virtualization.

Management of Enterprise Networks Today

Many enterprise networks consist of multiple layers of switches, often with VLANs to partition users into different administrative domains (for example, sales separated from engineering). Connected to these switches is a diverse range of appliances and devices⁹ required to manage the network and

implement security policy, sometimes requiring complex and fragile forwarding policies to put them in the path of packets. NonSDN switches are not programmable (by definition), so their forwarding and security policy is defined by what each vendor's proprietary configuration language provides. In some cases, an external system, such as an intrusion detection system (IDS), can make coarse changes to a network to implement dynamic security policy (for example, disable a host port if the IDS determines that a host on that port has been infected with malware).

Today's network operators are responsible for administering and integrating that wide range of appliances and devices, and it may be difficult to implement a specific security policy if the available devices do not have the necessary features to achieve the desired policy result. To operate and maintain a secure network with inflexible tools requires considerable skill and effort, and since the network cannot be programmed, opportunities to automate are limited. This is especially true when vendors either do not provide programmability or provide only proprietary automation technologies that operate best on a particular vendor's equipment.

Deploying Faucet in Today's Enterprise Networks

To realize the benefit of SDN you have to be able to deploy it. Deployment has to be easy and, ideally, incremental. To fit these real-life requirements, Faucet was designed to replace a conventional non-SDN switch, one for one, as shown in Figure 1, realizing the benefits of SDN in that network without necessitating notable infrastructure changes.

Faucet is deployed as a unit of two systems: a controller host (typically a host running Ubuntu Linux, running the Faucet controller application) and an OpenFlow switch (for example, an Allied Telesis x930 series switch), which are directly connected. The controller takes one configuration file, which describes which ports are in which VLANs (or if a port is in a VLAN trunk). The entire installation process, including creation of the configuration file (which resembles a nonSDN switch configuration file), has been reported to take only minutes in recent deployments.



While SDN as a technology continues to evolve and become even more programmable, Faucet and OpenFlow 1.3 hardware together are sufficient to realize benefits today.



Stopping at simple deployment, however, won't realize meaningful benefits. In fact, replacing a single switch with a controller machine (note that one controller machine can control many switches) *and* a switch takes more space and power. The benefits come from having the controller and control software both separate from the switch hardware, and entirely within the network operator's control, rather than being a closed vendor-provided appliance that cannot be reprogrammed.

Controllers can be numerous for high availability—Faucet supports redundant controllers—and as big or as small as required. Faucet is run in production at one site using a Raspberry Pi as a controller, which is practical because with Faucet the switch hardware does the forwarding—the controller does not have to be very powerful, because it does not have to forward traffic, leaving that to the higher-performance switch hardware.

Push Code to the Network on Green

"Push on green" refers to a philosophy of being able to create and test software in an automated fashion such that it is easy to detect a "green" condition of code—ready to roll out with as few bugs as possible.⁶ Google has published some of its strategies for deploying and managing large reliable software systems in the recent book, *Site Reliability Engineering*.³ SDN promises to help apply these strategies to networking software.

In keeping with this, the Faucet software stack includes a unit-testing framework,² so that new features can be developed with unit tests, and tests can be run against both simulated (Mininet virtual network) and hardware switches. Tests detect "green," and the operator can "push on green" with confidence that the system will work as tested. This accomplishes both feature-level and system-level testing—it is possible to catch system and integration problems at development time, well before deployment or even lab testing.

As an example, many non-SDN switches implement unicast flooding as part of the learning process,⁴ so that the switch can discover which hosts are connected to which ports. Learning works in this way: a host sends an Ether-

net packet with a destination unknown to the switch, and the switch floods the packet to all ports in the hope that the intended destination host will respond. This may not be desirable for security reasons, and in many non-SDN switches this behavior is hard-coded.

A Faucet feature that implements switch learning exclusively from Address Resolution Protocol (ARP) and neighbor discovery packets relieves Faucet from the need to flood unicast packets. The feature was implemented, tested with unit tests that included hardware and software, and pushed to github. On the controller machine, the operator ran “git pull” and then “service faucet restart,” accomplishing a complete controller upgrade and restart. Forwarding was interrupted for less than one second. Other features involving changes to the controller, even more ambitious features, can be deployed the same way.

As shown in Figure 2, Faucet implements all its features using OpenFlow 1.3 and multiple tables, and implements VLAN switching, IPv4 and IPv6 routing (both static and via the BGP routing protocol), access control lists (ACLs), port mirroring, and policy-based forwarding. The switch does all the forwarding, and no “hybrid” mode functionality is used on the switch. Hybrid mode is where a switch uses a mixture of nonprogrammable, non-SDN local processing, together with OpenFlow control. Faucet does not need such switch-local processing, and in our experience hybrid mode increases complexity and limits programmability by introducing the possibility of conflict between local and OpenFlow control.

A tiny fraction of traffic is copied to the controller so it can learn which hosts are on which ports and so the controller itself can resolve next hops (if routing is configured). The controller is generally idle unless hosts are added or move between ports (in which case the controller reprograms the pipeline as appropriate). Faucet has basic protection against control-plane attacks (for example, limiting spoofed Ethernet MAC addresses). Because the pipeline is entirely programmed by the controller, the network operator is free to make arbitrary changes to forwarding behavior by changing the controller software.



Faucet has basic protection against control-plane attacks. Because the pipeline is entirely programmed by the controller, the network operator is free to make arbitrary changes to forwarding behavior by changing the controller software.



When the hardware switch boots, it establishes an OpenFlow connection to the controller. The controller provisions the initial pipeline, including expectations for VLAN tags and any ACLs if configured, and adds “default-deny” rules (all unknown traffic is explicitly dropped). When a new host is detected, the switch sends a copy of the Ethernet header to the controller and (if unicast flooding is enabled) floods it to all other ports on the same VLAN or (if unicast flooding is disabled) floods it only if the packet is an IPv6 neighbor discovery or ARP packet. The controller then programs flows to cause future packets from this Ethernet source address to be forwarded by the switch. These flows periodically time out and are refreshed by the controller as necessary, which allows the switch to conserve resources.

Role of the CPN and Security of the Control Plane

The control-plane network (CPN) connects the controller machine and the switch on a dedicated port. In many deployments this is simply a good-quality three-foot Ethernet cable, which has been observed in production to be no less reliable than the internal connection between the CPU and data plane in a non-SDN switch. Indeed, should that three-foot cable fail, it can easily be replaced. In larger deployments where one controller machine controls several OpenFlow switches, the three-foot cable can be replaced by several Ethernet cables and a good-quality non-SDN switch. The controller connection between the switch and the Faucet controller can be secured with certificates or even MACsec-capable interfaces. (MACsec is the IEEE 802.1AE MAC security standard.)

Many switches allow configuration of the handling of the control connection being lost: “fail secure” (keep forwarding and using currently programmed flows until they expire) or “fail standalone” (revert to being essentially a nonprogrammable switch). Faucet implements expiry times on all flows, which causes forwarding to cease if no controller can be reached for a configurable period, so Faucet expects the switch to be in “fail secure” mode. It is possible to replace the CPN (and switch), and this has been done in production without interrupting

forwarding if done within the flow expiry time. (It is generally not possible, on the other hand, to replace the back plane in a standalone non-SDN switch without interrupting forwarding.)

Powerful Controllers Are Opportunities

In a non-SDN switch the embedded CPU is generally power- and cost-optimized. With Faucet, however, the controller can be a general-purpose computer or, indeed, a powerful server-class computer. This represents an opportunity to use the controller machine hardware as an open source network coprocessor or an alternative to a firewall appliance—the Faucet switch effectively provides lots of extra ports to a powerful machine.

In the first author's own deployment¹ the controller machine also runs Ubuntu's *ufw* (uncomplicated firewall) package, which implements NAT (network address translation) and firewalling; *Bro* (<https://www.bro.org/>), which is an IDS (intrusion detection system); and Internet Software Consortium's *dhcpd* (Dynamic Host Configuration Protocol daemon) to assign addresses dynamically. All three VLANs are trunked to a dedicated port on the controller machine. Faucet has ACLs on the host ports, which prevent hosts from spoofing the controller's IP addresses in each VLAN (so that proxy ARP attacks are not possible). Faucet ACLs operate across layers, so it is possible for an ACL entry to match, for example, Ethernet type, as well as IP address and MAC address.

More complex configurations are possible; for example, it would be possible, via the controller's trunk port, to assign a Linux container to each VLAN and run a separate *iptables* chain for every switch port. This achieves complete isolation and complete security policy customization on a port-by-port basis, without requiring any changes to the switch.

Faucet departs from common current network management practice in that it does not implement Simple Network Management Protocol (SNMP). Instead, the Faucet controller pushes statistics (bytes, packets, in and out from each port) to an external system. Faucet supports InfluxDB (<https://influxdata.com/>), which is an open source time series database. Using Faucet together with an open source

visualization system, Grafana (<http://grafana.org/>), it is possible to construct dashboards and run queries on current and historical data, as shown in Figure 3. Faucet can also produce a JSON (JavaScript Object Notation) log of statistics that could be translated and input into another system.

Northbound API. The SDN community continues to debate APIs that business, security, or other applications external to the controller should use to control the controller (for example, to ask the controller to prioritize a given user's traffic on the network). Faucet does not have a "one true" northbound API, because a generic API is not required. An operator can develop an application on top of Faucet that delivers Faucet a new configuration file and asks Faucet to apply it (for example, to change a user VLAN). Or, an operator might directly modify the controller code to add the desired feature, or add some other API convenient to the operator's needs (for example, to integrate the *Bro* IDS).

Relationship to Other SDN Controller Projects

The SDN community has several controller projects. Two well-known ones are ODL (OpenDaylight, <https://www.opendaylight.org/>) and ONOS (Open Network Operating System, <http://onosproject.org/>). Both controllers have many ambitions and have already served as technology demonstrators. As operational experience with SDN in the wider industry is still in its early stages, however, it is important to provide low-risk, incremental migration paths between today's networks and those aimed for by ODL and ONOS. Faucet could provide one such path and help inform the ongoing design of new network abstractions and programming frameworks. Furthermore, today's network operator community has traditionally not written its own software. While DevOps may be mainstream for many service operators, it is not for network operators, and it is important to make the benefits of such practices directly relevant to them.

Conclusion

The benefits of SDN have been difficult to realize because of a lack of software that is accessible to today's network operator community. While still a very

simple system, Faucet could be useful enough to operators that they may take the next step toward migrating to SDN, enabling them to adopt and enjoy the specific benefits of the rapid feature development, deployment, and testing Faucet provides.

Downloading Faucet. Faucet's source code for development can be found at <https://github.com/REANNZ/faucet>, or can be downloaded at <https://pypi.python.org/pypi/ryu-faucet/> (including Docker images) for easiest installation. C

Related articles on queue.acm.org

The Road to SDN

Nick Feamster et al.

<http://queue.acm.org/detail.cfm?id=2560327>

OpenFlow: A Radical New Idea in Networking

Thomas A. Limoncelli

<http://queue.acm.org/detail.cfm?id=2305856>

A Purpose-built Global Network: Google's Move to SDN

<http://queue.acm.org/detail.cfm?id=2856460>

References

1. Bailey, J. NFV/firewall offload with Faucet. Faucet SDN, 2016; <http://faucet-sdn.blogspot.co.nz/2016/05/nfv-firewall-offload-with-faucet.html>.
2. Bailey, J. Unit-testing framework. Faucet SDN, 2016; <http://faucet-sdn.blogspot.co.nz/2016/06/unit-testing-framework.html>.
3. Beyer, B., Jones, C., Petofo, J. and Murphy, N.R., eds. *Site Reliability Engineering*. O'Reilly Media, 2016.
4. Cisco. Unicast flooding in switched campus networks, 2016; <http://www.cisco.com/c/en/us/support/docs/switches/catalyst-6000-series-switches/23563-143.html>.
5. Faucet; <https://github.com/REANNZ/faucet>.
6. Klein, D.V., Betser, D.M. and Monroe, M.G. Making 'push on green' a reality: Issues and actions involved in maintaining a production service. Research at Google, 2014; <http://research.google.com/pubs/pub42576.html>.
7. McKeown, N. et al. OpenFlow: Enabling innovation in campus networks. *ACM SIGCOMM Computer Communication Review* 38, 2 (2008), 69–74.
8. Open Networking Foundation. OpenFlow Switch Specification, ver. 1.3.3, 2013; <https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-spec-v1.3.3.pdf>.
9. Sherry, J. and Ratnasamy, S. A survey of enterprise middlebox deployments. Technical Report UCB/EECS-2012-24. University of California at Berkeley; <http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-24.pdf>.
10. Vahdat, A. Pulling back the curtain on Google's network infrastructure. Google Research Blog, 2015; <http://googleresearch.blogspot.com/2015/08/pulling-back-curtain-on-googles-network.html>.

Josh Bailey is a staff software engineer at Google, working on network management and research projects for the past 11 years. He is based in Wellington, New Zealand.

Stephen Stuart is a distinguished software engineer at Google for the past 13 years, based in Mountain View, CA.

Copyright held by owners/authors.

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

Research for Practice: Web Security and Mobile Web Computing

OUR THIRD INSTALLMENT of Research for Practice brings readings spanning programming languages, compilers, privacy, and the mobile Web.

First, **Jean Yang** provides an overview of how to use information flow techniques to build programs that are secure by construction. As Yang writes, information flow is a conceptually simple “clean idea”: the flow of sensitive information across program variables and control statements can be tracked to determine whether information may in fact leak. Making information flow practical is a major challenge, however. Instead of relying on programmers to track information flow, how can compilers and language runtimes be made to do the heavy lifting? How can application writers easily express their privacy policies and understand the implications of a given policy for the set of values that an application user may see? Yang’s set of papers directly addresses these questions

via a clever mix of techniques from compilers, systems, and language design. This focus on theory made practical is an excellent topic for RfP.

Second, **Vijay Janapa Reddi** and **Yuhao Zhu** provide an overview of the challenges for the future of the mobile Web. Mobile represents a major frontier in personal computing, with extreme growth in adoption and data volume. Accordingly, Reddi and Zhu outline three major ongoing challenges in mobile Web computing: responsiveness of resource loading, energy efficiency of computing devices, and making efficient use of data. In their citations, Reddi and Zhu draw on a set of techniques spanning browsers, programming languages, and data proxying to illustrate the opportunity for “cross-layer optimization” in addressing these challenges. Specifically, by redesigning core components of the Web stack, such as caches and resource-fetching logic, systems operators can improve users’ mobile Web experience. This opportunity for co-design is not simply theoretical: Reddi and Zhu’s third citation describes a mobile-optimized compression proxy that is already running in production at Google.

As always, our goal in RfP is to allow readers to become experts in the latest, practically oriented topics in computer science research in a weekend afternoon’s worth of reading time. I am grateful to this installment’s experts for generously contributing such a strong set of contributions, and, as always, we welcome your feedback!

— *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.

» 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.



Practical Information Flow for Web Security

By Jean Yang

Information leaks have become so common that many have given up hope when it comes to information security.³ Data breaches are inevitable anyway, some say.¹ I don't even go on the Internet anymore, other (computers) say.⁶

This despair has led yet others to the Last Resort: Reasoning about what our programs actually do. For years, bugs didn't matter as long as your robot could sing. If your program can go twice the speed it did yesterday, who cares what outputs it gives you? But we are starting to learn the hard way that no amount of razzle-dazzle can make up for Facebook leaking your phone number to the people you didn't invite to the party.⁴

This realization is leading us to a new age, one in which reasoning techniques that previously seemed unnecessarily baroque are coming into fashion. Growing pressure from regulators is finally making it increasingly popular to use precise program analysis to ensure software security.⁵ Growing demand for producing Web applications quickly makes it relevant to develop new paradigms—well-specified ones, at that—for creating secure-by-construction software.

The construction of secure software means solving the important problem of *information flow*. Most of us have heard of trapdoor ways to access information we should not see. For example, one researcher showed that it is possible to discover the phone numbers of thousands of Facebook users simply by searching for random phone numbers.² Many such leaks occur not because a system shows sensitive values directly, but because it shows the results of computations—such as search—on sensitive values. Preventing these leaks requires implementing policies not only on sensitive values themselves, but also whenever computations may be affected by sensitive values.

Enforcing policies correctly with respect to information flow means reasoning about sensitive values and policies as they flow through increasingly complex programs, making sure to reveal only information consistent with the privileges associ-

ated with each user. There is a body of work dedicated to compile-time and runtime techniques for tracking values through programs for ensuring correct information flow.

While information flow is a clean idea, getting it to work on real programs and systems requires solving many hard problems. The three papers presented here focus on solving the problem of secure information flow for Web applications. The first one describes an approach for taking trust out of Web applications and shifting it instead to the framework and compiler. The second describes a fully dynamic enforcement technique implemented in a Web framework that requires programmers to specify each policy only once. The third describes a Web framework that customizes program behavior based on the policies and viewing context.

Shifting Trust to the Framework and Compiler through Language-Based Enforcement

Chong, S., Vikram, K. and Myers, A.C.

SIF: Enforcing confidentiality and integrity in Web applications. *Proceedings of the 16th Usenix Security Symposium, 2007.*

<https://www.usenix.org/conference/16th-usenix-security-symposium/sif-enforcing-confidentiality-and-integrity-Web>

In securing Web applications, a major source of the burden on programmers involves reasoning about how information may be leaked through computations across different parts of an application and across requests. Without additional checks and balances, the programmer must be fully trusted to do this correctly.

This first selection presents a framework that shifts trust from the application to the framework and compiler. The Servlet Information Flow (SIF) framework follows a line of work in language-based information flow focused on checking programs against specifications of security policies. Built using the Java servlet framework, SIF prevents many common sources of information flow—for example, those across multiple requests. SIF applications are written in Jif, a language that extends Java with programmer-provided labels specifying policies for information flow. SIF uses a combina-

tion of compile-time and runtime enforcement to ensure security policies are enforced from the time a request is submitted to when it is returned, with modest enforcement overhead. The major contribution of the SIF work is in showing how to provide assurance (much of it at compile time) about information flow guarantees in complex, dynamic Web applications.

Mitigating Annotation Burden through Principled Containment

Giffin, D.B. et al.

Hails: Protecting data privacy in untrusted Web applications. *Proceedings of the 10th Usenix Symposium on Operating Systems Design and Implementation, 2012.*

<https://www.usenix.org/node/170829>

While compile-time checking approaches are great for providing assurance about program security, they often require nontrivial programmer effort. The programmer must not only correctly construct programs with respect to information flow, but also annotate the program with the desired policies.

An alternative approach is confinement: running untrusted code in a restricted way to prevent the code from exhibiting undesired behavior. For information flow, confinement takes the form of tagging sensitive values, tracking them through computations, and checking tags at application endpoints. Such dynamic approaches are often more popular because they require little input from the programmer.

This paper presents Hails, a Web framework for principled containment. Hails extends the standard MVC (model-view-controller) paradigm to include policies, implementing the MPVC (model-policy-view-controller) paradigm where the programmer may specify label-based policies separately from the rest of the program. Built in Haskell, Hails uses the LIO (labeled IO) library to enforce security policies at the thread/context level and MAC (mandatory access control) to mediate access to resources such as the database. It has good performance for an information flow control framework, handling approximately 47.8K requests per second.

Hails has been used to build several Web applications, and the startup Intrinsic is using a commercial version of Hails.

The Hails work shows it is possible to enforce information flow in Web applications with negligible overhead, without requiring programmers to change how they have been programming.

Shifting Implementation Burden to the Framework

Yang, J., et al.

Precise, dynamic information flow for database-backed applications. *Proceedings of the 37th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2016, 631–647. <http://dl.acm.org/citation.cfm?id=2908098>

With the previous two approaches, the programmer remains burdened with constructing programs correctly with respect to information flow. Without a change in the underlying execution model, the most any framework can do is raise exceptions or silently fail when policies are violated.

This paper looks at what the Web programming model might look like if information flow policies could be factored out of programs the way memory-managed languages factor out allocation and deallocation. The paper presents Jacqueline, an MPVC framework that allows programmers to specify how to compute an alternative default for each data value; and high-level policies about when to show each value that may contain database queries and/or depend on sensitive values.

A plausible default for a sensitive location value is the corresponding city. A valid policy is allowing a viewer to see the location only if the viewer is within some radius of the location. This paper presents an implementation strategy for Jacqueline that works with existing SQL databases. While the paper focuses more on demonstrating feasibility than on the nuts and bolts of Web security, it de-risks the approach for practitioners who may want to adopt it.

Final Thoughts

The past few years have seen a gradual movement toward the adoption of practical information flow: first with containment, then with microcontainers and microsegmentation. These techniques control which devices and services can interact with policies for software-defined infrastructures such as iptables and software-defined network-

ing. Illumio, vArmour, and GuardiCore are three among the many startups in the microsegmentation space. This evolution toward finer-grained approaches shows that people are becoming more open to the system re-architecting and runtime overheads that come with information flow control approaches. As security becomes even more important and information flow techniques become more practical, the shift toward more adoption will continue.

Acknowledgments. Thanks to A. Aulich, S. Chong, V. Iozzo, L. Meyerovich, and D. Stefan. 

References

1. Balluck, K. Corporate data breaches 'inevitable,' expert says. *The Hill* (Nov. 30 2014); <http://thehill.com/policy/cybersecurity/225550-cybersecurity-expert-data-breaches-inevitable>.
2. Cunningham, M. Facebook security flaw could leak your personal info to criminals. *Komando.com* (Aug. 10, 2015); <http://bit.ly/2fRXp8L>.
3. Information is beautiful. World's biggest data breaches, 2016; <http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/>.
4. Gellman, B. and Poitras, L. U.S., British intelligence mining data from nine U.S. Internet companies in broad, secret program. *Washington Post* (June 7, 2013); <http://wapo.st/1LcAw6p>.
5. Open Web Application Security Project (OWASP). Static code analysis, 2016; https://www.owasp.org/index.php/Static_Code_Analysis.
6. Zetter, K. Hacker lexicon: What is an air gap? *Wired* (Dec. 8, 2014); <http://www.wired.com/2014/12/hacker-lexicon-air-gap/>.

Jean Yang is an assistant professor in the computer science department at Carnegie Mellon University. In 2015 she cofounded the Cybersecurity Factory accelerator to bridge the gap between research and practice in cybersecurity.



The Red Future of Mobile Web Computing
By Vijay Janapa Reddi and Yuhao Zhu

The Web is on the cusp of a new evolution, driven by today's most pervasive personal computing platform—mobile devices. At present, there are more than three billion Web-connected mobile devices. By 2020, there will be 50 billion such devices. In many markets around the world mobile Web traffic volume exceeds desktop Web traffic, and it continues to grow in double digits.

Three significant challenges stand in the way of the future mobile Web. The papers selected here focus on carefully addressing these challenges. The first

major challenge is the *responsiveness* of Web applications. It is estimated that a one-second delay in Web page load time costs Amazon \$1.6 billion in annual sales lost, since mobile users abandon a Web service altogether if the Web page takes too long to load. Google loses eight million searches from a four-tenths-of-a-second slowdown in search-results generation. A key bottleneck of mobile Web responsiveness is resource loading. The number of objects in today's Web pages is already on the order of hundreds, and it continues to grow steadily. Future mobile Web computing systems must improve resource-loading performance, which is the focus of the first paper.

The second major challenge is *energy efficiency*. Mobile devices are severely constrained by the battery. While computing capability driven by Moore's Law advances approximately every two years, battery capacity doubles every 10 years—creating a widening gap between computational horsepower and the energy needed to power the device. Therefore, future mobile Web computing must be energy efficient. The second paper in our selection proposes Web programming language support for energy efficiency.

The third major challenge is *data usage*. A significant amount of future mobile Web usage will come from emerging markets in developing countries where the cost of mobile data is prohibitively large. To accelerate the Web's growth in emerging markets, future mobile Web computing infrastructure must serve data consciously. The final paper discusses how to design a practical and efficient HTTP data compression proxy service that operates at Google's scale.

Developers and system architects must optimize for RED (responsiveness, energy efficiency, and data usage), ideally together, to usher in a new generation of mobile Web computing.

Intelligent Resource Loading For Responsiveness

Netravali et al.

Polaris: Faster page loads using fine-grained dependency tracking. *Proceedings of the 13th Usenix Symposium on Networked Systems Design and Implementation*, 2016. <https://www.usenix.org/conference/nsdi16/technical-sessions/presentation/netravali>

A key bottleneck for mobile Web responsiveness is resource loading. The

bottleneck stems from the increasing number of objects (for example, images and Cascading Style Sheets files) on a Web page. According to the HTTP Archive, over the past three years alone, Web pages have doubled in size. Therefore, improving resource-loading performance is crucial for improving the overall mobile Web experience.

Resource loading is largely determined by the critical path of the resources that Web browsers load to render a page. This critical path, in the form of a resource-dependency graph, is not revealed to Web browsers statically. Therefore, today's browsers make conservative decisions during resource loading. To avoid resource-dependency violations, a Web browser typically constrains its resource-loading concurrency, which results in reduced performance.

Polaris is a system for speeding up the loading of Web page resources, an important step in coping with the surge in mobile Web resources. Polaris constructs a precise resource-dependency graph offline, and it uses the graph at runtime to determine an optimal resource-loading schedule. The resulting schedule maximizes concurrency and, therefore, drastically improves mobile Web performance. Polaris also stands out because of its transparent design. It runs on top of unmodified Web browsers without the intervention of either Web application or browser developers. Such a design minimizes the deployment inconvenience and increases its chances of adoption, two factors that are essential for deploying the Web effectively.

Web Language Support for Energy Efficiency

Zhu, Y., Reddi, J.

GreenWeb: Language extensions for energy-efficient mobile Web computing.
Proceedings of the 37th ACM SIGPLAN Conference on Programming Language Design and Implementation, 2016, 145–160.
<http://dl.acm.org/citation.cfm?id=2908082>

Energy efficiency is the single most critical constraint on mobile devices that lack an external power supply. Web runtimes (typically the browser engine) must start to budget Web application energy usage wisely, informed by user QoS constraints. End-user QoS information, however, is largely unaccounted for in today's Web programming languages.

The philosophy behind GreenWeb is that application developers provide minimal yet vital QoS information to guide the browser's runtime energy optimizations. Empowering a new generation of energy-conscious Web application developers necessitates new programming abstractions at the language level. GreenWeb proposes two new language constructs, *QoS type* and *QoS target*, to capture the critical aspects of user QoS experience. With the developer-assisted QoS information, a GreenWeb browser determines how to deliver the specified user QoS expectation while minimizing the device's energy consumption.

GreenWeb does not enforce any particular runtime implementation. As an example, the authors demonstrate one implementation using ACMP (asymmetric chip-multiprocessor) hardware. ACMP is an energy-efficient heterogeneous architecture that mobile hardware vendors such as ARM, Samsung, and Qualcomm have widely adopted—you probably have one in your pocket. Leveraging the language annotations as hints, the GreenWeb browser dynamically schedules execution on the ACMP hardware to achieve energy savings and prolong battery life.

Data Consciousness in Emerging Markets

Agababov, V. et al.

Flywheel: Google's data compression proxy for the mobile Web. *Proceedings of the 12th Usenix Symposium on Networked Systems Design and Implementation*, 2015;
<http://research.google.com/pubs/pub43447.html>

The mobile Web is crucial in emerging markets. The first order of impedance for the mobile Web in emerging markets is the high cost of data, more so than performance or energy efficiency. It is not uncommon for spending on mobile data to be more than half of an individual's income in developing countries. Therefore, reducing the amount of data transmitted is essential.

Flywheel from Google is a compression proxy system to make the mobile Web conscious of data usage. Compression proxies to reduce data usage (and to improve latency) are not a new idea. Flywheel, however, demonstrates that while the core of the proxy server is compression,

“At present there are more than three billion Web-connected mobile devices. By 2020, there will be 50 billion such devices.”

there are many design concerns to consider that demand a significant amount of engineering effort, especially to make such a system practical at Google scale. Examples of the design concerns include fault tolerance and availability upon request anomalies, safe browsing, robustness against middlebox optimizations, and so on. Moreover, drawing from large-scale measurement results, the authors present interesting performance results that might not have been observable from small-scale experiments. For example, the impact of data compression on latency reduction is highly dependent on the user population, metric of interest, and Web page characteristics.

Conclusion

We advocate addressing the RED challenge holistically. This will entail optimizations that span the different system layers synergistically. The three papers in our selection are a first step toward such cross-layer optimization efforts. With additional synergy we will likely uncover more room for optimization than if each of the layers worked in isolation. It is time that we as a community make the Web great again in the emerging era. 

Vijay Janapa Reddi is an assistant professor in the Department of Electrical and Computer Engineering at the University of Texas at Austin.

Yuhao Zhu is a Ph.D. candidate at the University of Texas at Austin.

Copyright held by owners/authors.
Publication rights licensed to ACM. \$15.00

DOI:10.1145/2976758

Moore's Law is one small component in an exponentially growing planetary computing ecosystem.

BY PETER J. DENNING AND TED G. LEWIS

Exponential Laws of Computing Growth

IN A FORECASTING exercise, Gordon Earle Moore, co-founder of Intel, plotted data on the number of components—transistors, resistors, and capacitors—in chips made from 1959 to 1965. He saw an approximate straight line on log paper (see Figure 1). Extrapolating the line, he speculated that the number of components would grow from 2^6 in 1965 to 2^{16} in 1975, doubling every year. His 1965–1975 forecast came true. In 1975, with more data, he revised the estimate of the doubling period to two years. In those days, doubling components also doubled chip speed because the greater number of components could perform more powerful operations and smaller circuits allowed faster clock speeds. Later, Moore's Intel colleague David House claimed the doubling time

for speed should be taken as 18 months because of increasing clock speed, whereas Moore maintained that the doubling time for components was 24 months. But clock speed stabilized around 2000 because faster speeds caused more heat dissipation than chips could withstand. Since then, the faster speeds are achieved with multi-core chips at the same clock frequency.

Moore's Law is one of the most durable technology forecasts ever made.^{10,20,31,33} It is the emblem of the information age, the relentless march of the computer chip enabling a technical, economic, and social revolution never before experienced by humanity.

The standard explanation for Moore's Law is that the law is not really a law at all, but only an empirical, self-fulfilling relationship driven by economic forces. This explanation is too weak, however, to explain why the law has worked for more than 50 years and why exponential growth works not only at the chip level but also at the system and market levels. Consider two prominent cases of systems evolution.

Supercomputers are complete systems, including massively parallel arrays of chips, interconnection networks, memory systems, caches, I/O systems, cooling systems, languages for expressing parallel computations, and compilers. Various groups have tracked these systems over the years. Figure 2 is a composite graph of data from these groups on the speeds of the fastest computers since 1940. The performance of these computers has grown exponentially. One of

» key insights

- Exponential growth seems to be unique to computing and information technologies and their markets, stimulating continued economic, social, and political disruptions.
- Exponential growth occurs simultaneously at all levels of the computing ecosystem—chips, systems, adopting communities.
- Technology jumping sustains exponential growth as companies switch to new technologies when the current ones reach their points of diminishing return.

ILLUSTRATION BY PETER GROWTHER ASSOCIATES

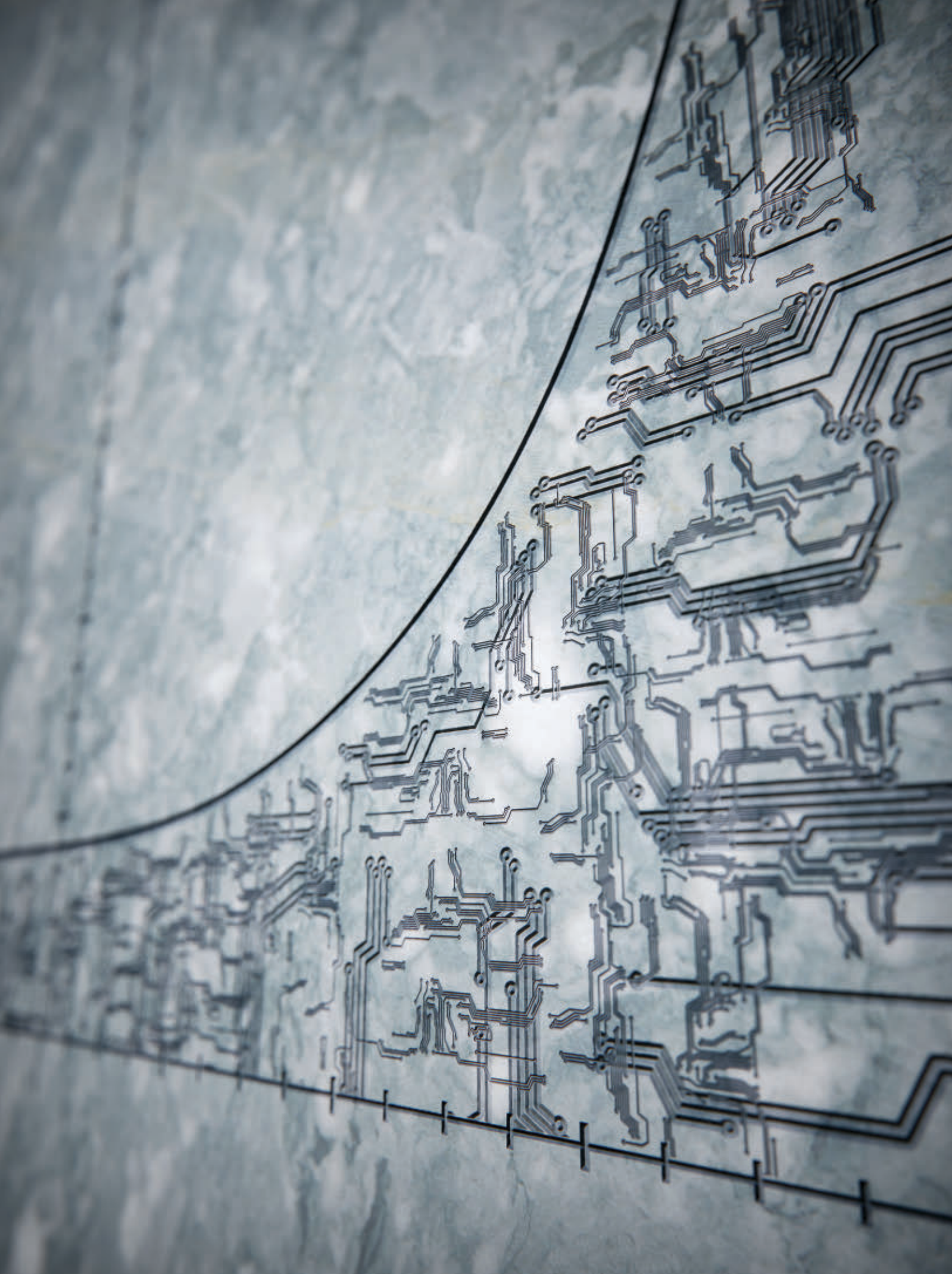
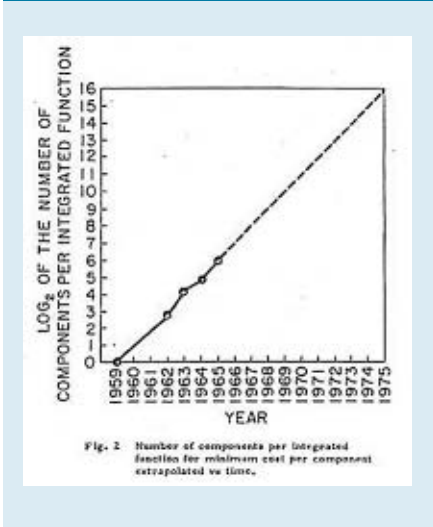


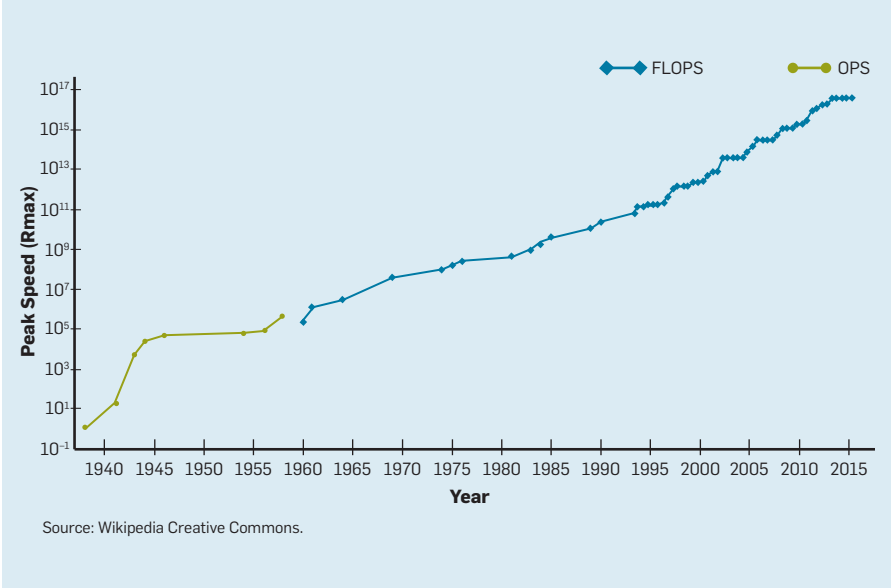
Figure 1. Moore's original prediction graph showed component count followed a straight line when plotted on log paper.²⁶



the tracking groups, TOP500 (<https://www.top500.org/>), has used a Linpack benchmark since 1993 to compare the fastest machines at each point in time for mathematical software, noting that the growth rate may be slowing because the market for such machines is slowing. The speeds of supercomputing systems depend on at least eight technologies besides the chips. How do we explain the exponential growth in this case?

Ray Kurzweil, futurist and author of *The Singularity Is Near: When Humans Transcend Biology*, formulated a set of predictions about information technology by constructing a graph of the computational speed growth over five generations of information technol-

Figure 2. Speeds of the fastest computers from 1940 show an exponential rise in speed. From 1965 to 2015, the growth was a factor of 12 orders of 10 over 50 years, or a doubling approximately every 1.3 years.



ogy (see Figure 3).²⁵ He projected that this remarkable exponential doubling trend would continue for another 100 years, relying on jumps to new technologies every couple of decades. He also forecast a controversial claim—a “singularity” around 2040 when artificial intelligence will exceed human intelligence.^{25,32} The exponential growth in this case clearly does not depend on Moore’s Law at all. How do we explain the exponential growth in this case?

The three kinds of exponential growth, as noted—doubling of components, speeds, and technology adoptions—have all been lumped under the heading of Moore’s Law. Because the

original Moore’s Law applies only to components on chips, not to systems or families of technologies, other phenomena must be at work. We will use the term “Moore’s Law” for the component-doubling rule Moore proposed and “exponential growth” for all the other performance measures that plot as straight lines on log paper. What drives the exponential growth effect? Can we continue to expect exponential growth in the computational power of our technologies?

Exponential growth depends on three levels of adoption in the computing ecosystem (see the table here). The chip level is the domain of Moore’s Law, as noted. However, the faster chips cannot realize their potential unless the host computer system supports the faster speeds and unless application workloads provide sufficient parallel computational work to keep the chips busy. And the faster systems cannot reach their potential without rapid adoption by the user community. The improvement process at all three levels must be exponential; otherwise, the system or community level would be a bottleneck, and we would not observe the effects often described as Moore’s Law.

With supporting mathematical models, we will show what enables exponential doubling at each level. Information technology may be unique in being able to sustain exponential growth at

Three levels of exponential growth in the computing ecosystem.	
Level	Explanation
Chip	Chip designers found technology paths for reducing component dimensions using Dennard Scaling ⁶ until around 2000, when heat-dissipation problems prevented clocks faster than about 3GHz. Since then, they have considered a host of methods, including multicore and clock distribution, to reduce power consumption and keep the components busy.
System	Improvements in chips, parallelism, cache, memory interconnects, networks, languages, compilers, and cooling enable a computer system to periodically double its speed and relieve performance bottlenecks in the system. Data-intensive workloads present a wealth of parallel threads sufficient to keep any multicore system busy.
Community	New system generations are highly attractive innovation enablers, and their adoption spreads exponentially in user communities.

all three levels. We will conclude that Moore's Law and exponential doubling have scientific bases. Moreover, the exponential doubling process is likely to continue across multiple technologies for decades to come.

Self-Fulfillment

The continuing achievement signified by Moore's Law is critically important to the digital economy. Economist Richard G. Anderson said, "Numerous studies have traced the cause of the productivity acceleration to technological innovations in the production of semiconductors that sharply reduced the prices of such components and of the products that contain them (as well as expanding the capabilities of such products)."¹ Robert Colwell, Director of DARPA's Microsystems Technology Office, echoes the same conclusion, which is why DARPA has invested in overcoming technology bottlenecks in post-Moore's-Law technologies.⁵ If and when Moore's Law ends, that end's impact on the economy will be profound.

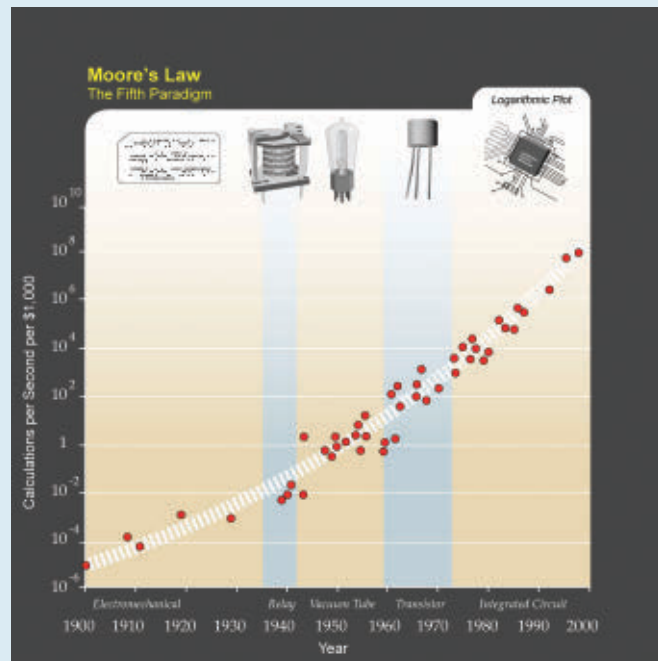
It is no wonder then that the standard explanation of the law is economic; it became a self-fulfilling prophecy of all chip companies to push the technology to meet the expected exponential growth and sustain their markets. A self-fulfilling prophecy is a prediction that causes itself to become true. For most of the past 50-plus years of computing, designers have emphasized performance. Faster is better. To achieve greater speed, chip architects increased component density by adding more registers, higher-level functions, cache memory, and multiple cores to the same chip area and the same power dissipation. Moore's Law became a design objective.

Designers optimize placement of components on highly constrained real estate, seeking to fill every square nanometer of area. Doubling component density—and therefore the number of components per nanometer of area—was not an outrageous objective because it requires only a reduction of 30% in both dimensions of 2D components. To achieve the next generation of Moore's Law, designers halve the area of each component, which means reducing each dimension to $\sqrt{1/2} = 0.71$ of its former value; we call this the "square

root reduction rule." Figure 4 shows that the die size of chips has consistently followed this rule over many generations. Wu et al.³² extended a

quantum-limit argument offered by Nobel physicist Richard Feynman¹² in 1985 to conclude that this downward scaling process could continue until

Figure 3. Kurzweil's graph of speed of information technologies since 1900 spans five families of technologies. From 1900 to 2000, the growth was 14 orders of 10 over 100 years, or a doubling approximately every 1.3 years.



Source: <http://www.kurzweilai.net>

Figure 4. Logarithm of actual versus predicted feature size since 1970 matches a straight line with regression coefficient $R^2 = 0.97$. Future sizes are predicted by dividing the previous size by $\sqrt{2}$; see the open triangles and dotted line. Future sizes two generations into the future are close to half the current sizes; see the square dots.

Data from David Harris, Lecture 21: "Scaling and Economics," Harvey Mudd College, Claremont, CA, 2004 (<http://pages.hmc.edu/harris/class/e158/04/lect21.pdf>) and Zvi Or-Bach, "Is the Cost Reduction Associated with Scaling Over?" Monolith3D, San Jose, CA, 2004 (<http://www.monolith3d.com/blog/is-the-cost-reduction-associated-with-scaling-over>).

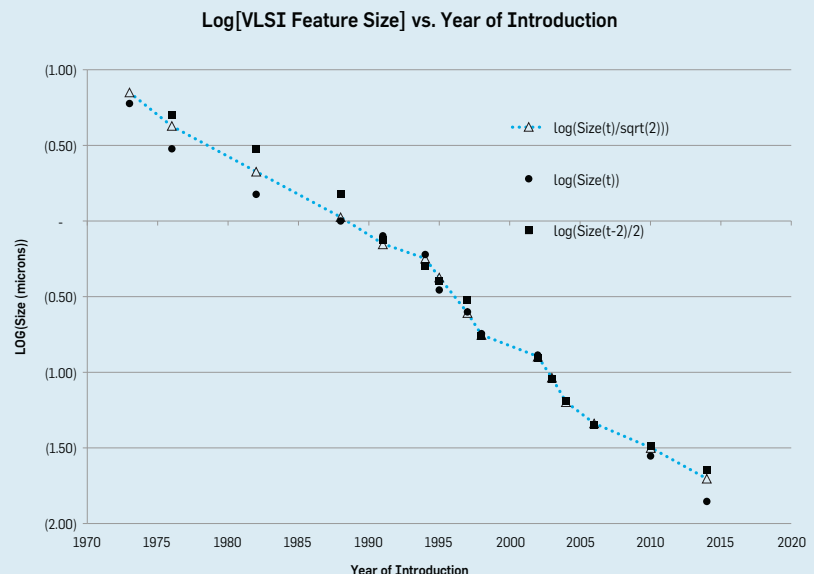
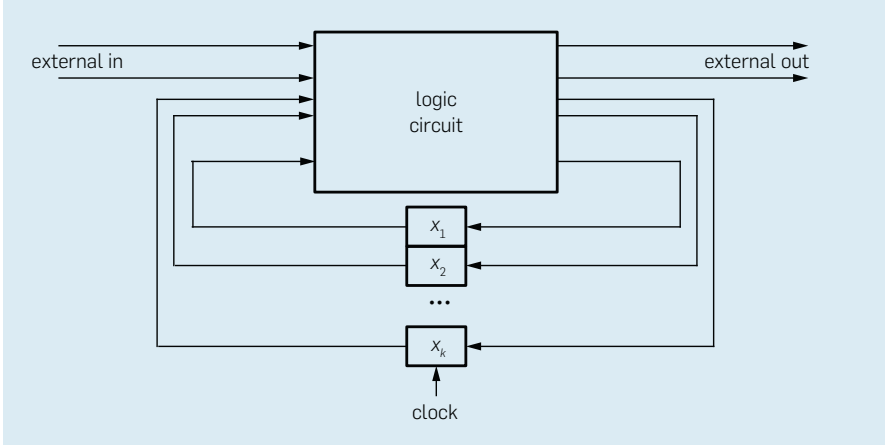


Figure 5. Computer subsystems are organized as stateless logic circuits (AND, OR, NOT gates without feedback loops) driving flip-flops that record the subsystem's state. Clock pulses trigger the flip-flops ($x_1, \dots, x_k$) to assume the states generated by the logic outputs. Too short a clock interval risks a metastable state because the logic outputs driving the flip-flops may not have settled since the last clock pulse.



device sizes approached the Compton wavelength of an electron, which will happen by approximately 2036.

Motivated by the promise of enormous economic payoffs, designers overcame many challenges to sustain this rule. Even so, we find self-fulfillment to be an unsatisfying explanation of the persistence of Moore's Law. Why does the law not work for other technologies? What if systems had too many bottlenecks or workloads contained insufficient parallelism? What if people failed to adopt new technologies?

Moore's Law at the Chip Level

What is special about information technologies that makes exponential growth a possibility? This possibility is not available for every technology. We might wish for automobiles that travel 1,600 miles on a gallon of gasoline—approximately six doublings (2^6) better than today's most efficient automobiles—but wishing will not make it happen. There is no technology path leading to that level of automobile efficiency. What is different about chip technology?

The answer is that the basic performance measure of computing systems is computational steps per unit time (or energy). Twice as many components enable twice as many computational steps. And the regularity of components enables doubling them by scaling down size.

A chip is made of a very large number of simple basic components, mostly transistors and interconnecting wires. Doubling the number of components

in a chip generation is a feasible goal, as in Figure 2, and a method called "Dennard scaling" (described in the following paragraphs) was invented to accomplish it. Few other technologies (such as automobiles) feature complex systems composed of large numbers of identical parts.

Component doubling every chip generation may be feasible but is not easy. The underlying silicon technology, CMOS, cannot support the continued component reductions. What are the problems? And what options are available to overcome them? We cover three main ones:

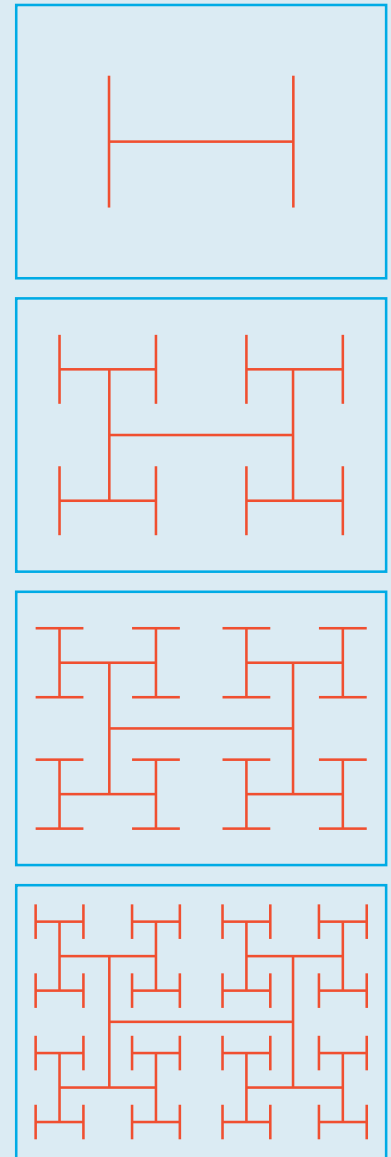
Path to the square root reduction rule. Dennard scaling, which defined a path to the square root reduction rule for nearly 30 years, why it came to an end, and what has been done to increase computational speed since then;

Clock trees and clock distribution. Why do we have clocks? What is the problem with clocked circuits? Could it be solved with an additional clock distribution layer on the chip? Or by replacing clocked circuits with fast asynchronous (unclocked) circuits?; and

Taxonomy. A taxonomy of issues needed to keep Moore's Law at chip level going for more generations.

Dennard Scaling. In 1974, electrical engineer and inventor Robert Dennard and his team at IBM proposed a method to scale down transistors while maintaining a constant power density.⁶ The power density is energy dissipated in a square unit of chip area. It is proportional to the switching speed and

Figure 6. Four iterations of the space-filling H-tree fractal show how to quadruple the number of terminal nodes by halving the size of each "H." Each iteration appends a half-size "H" to the terminal nodes of the previous "H." A clock signal is injected at the center and arrives at all the terminal nodes simultaneously because the H-tree is balanced; see <http://www.tamurajones.net/FractalGenealogy.xhtml>



the number of transistors in a unit area. Greater power densities mean more heat to dissipate—and too much heat will burn up the chip. Dennard scaling says that power density stays constant as transistors get smaller so the power used is proportional to area.

Reducing the size of a subsystem and its components can allow for the clock interval to be shortened because the

logic gates switch faster and the wires connecting them are shorter. As noted, however, there is a limit to this approach because the increased number of state switches produces more heat. Engineers were able to increase clock speeds from approximately 5MHz in 1981 (IBM PC) to approximately 2GHz in 2000, leveling off at approximately 3.5GHz since 2002 (Intel Pentium 4). The cost of heat-sink technology to support greater clock speeds is prohibitive.

Even when clock speeds were held constant, chip engineers started to discover in the 1990s that leakage and quantum-tunneling effects became significant at small dimensions and produced more heat than Dennard scaling predicted. Dennard scaling was no longer a reliable pathway to reducing component size.

Multicore architectures were the response to the demise of Dennard scaling. The first two-core chips had twice as many components as the previous one-core chips. They were organized as two CPUs running in parallel at the same limiting clock speed (approximately 3.5GHz). Two cores could achieve twice the one-core speed if the computational workload had multiple computational threads. However, doubling on-chip cores every generation has its own heat-dissipation problems that will limit how many cores can be usefully placed on a chip.¹¹ Moreover, multicore architecture pushed some of the responsibility for speedup to multi-threaded programming and parallelizing compilers. Multicore parallelism is a fine strategy but draws programmers into parallel programming, something many were never trained to do.²

Metastability, clocks, and clock trees. Clocks became an integral feature of computer logic circuits in the 1940s because engineers could quickly and simply avoid a host of timing-dependent failures that result from metastable behavior in computer circuits. If the clock interval is too short, the flip-flops recording a subsystem's state can be triggered before their inputs have settled down, risking internal oscillations that freeze circuits or cause other malfunctions (see Figure 5).⁷

Metastability is also an issue when asynchronous subsystems (no common clock) can sense unsettled signals in an attempt to synchronize with each

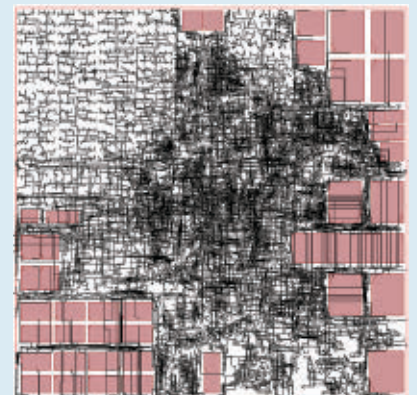
Few other technologies (such as automobiles) feature complex systems composed of large numbers of identical parts.

other. Designers use special protocols (such as ready-acknowledge signaling) to minimize the risk of synchronization failure in those cases.⁷

A major design problem is transmitting signals from a clock to all the components that need clock signals. The wires that do this take space on the chip, and there is a propagation delay along each wire. The term “clock skew” refers to the difference between the path with shortest delay and the path with longest delay. To avoid metastability, engineers must choose the clock time long enough to overcome skew.

The ideal circuit to distribute clock signals is a balanced tree with the clock at its root and the components at its leaves. In a balanced tree, all paths are the same length, and there is no skew. Many designers considered the H-tree fractal, a mathematical model of a geometric tree structure that replicates at ever-finer dimensions with each generation (see Figure 6). However, the H-fractal is space-filling, meaning that as the dimension of the tree gets larger the area occupied by the wires eventually fills the chip, leaving little room for active components. This forces chip designers to use more ad hoc methods that depend on unbalanced trees restricted to a portion of the area (see Figure 7). The unbalanced trees have a much larger number of leaf nodes than needed, so the designer can choose leaf nodes nearest the components needing signals and minimize clock skew. Some designers propose to put

Figure 7. Design of an actual chip has approximately half its area devoted to the clock tree (dense lines) and half to its actual components (colored boxes). Image courtesy of Sung Kyu Lim at The Georgia Institute of Technology.²⁷



Technology Jumping in Pursuit of Moore's Law

Wu et al.³³ conducted a study of possible directions for continuing the Moore's Law effect, citing seven major barriers to the continuation of the current Moore's Law with CMOS technology:

- Performance demand of the processor.* Exponential growth cannot be sustained inside a single technology;
- Power consumption and heat dissipation.* They grow worse per unit area as component size decreases;
- Communication costs of moving data through networks, interconnects, and caches.* They grow with the number of cores served;
- Tunneling effect.* Electrons jump narrow insulating barriers;
- Quantum limit to Moore's Law.* Compton wavelength is the fundamental limit to measuring the position of electrons; if component dimensions become that small, it will no longer be possible to tell where electrons are and whether they are being switched properly;
- Economic limit to Moore's Law.* Costs of R&D and manufacturing are rising exponentially, making it increasingly difficult for each next generation to be economically viable;
- On-board limits.* As designers move more functions onto a chip, the chip's performance depends on all the technologies, not just the logic circuits; for example, mixed-signal circuits (such as analog-digital converters and digital signal processors) are limited by sampling frequencies and sensitive to fabrication variations in transistors; and
- Mobile technologies.* Smartphones and multimedia phones present a phalanx of barriers to performance improvement, including increasing demand for bandwidth, concern for power reduction and battery life, limits on size and weight, and limits on what consumers are willing to pay.

Despite these barriers, Wu et al. saw eight ways new technology could address them:

- DNA scaffolding.* Employ DNA scaffolding technologies to build (grow) circuit boards.
- 3D fabrication.* Move to 3D fabrication;
- Carbon nanotubes and graphene.* Build components from carbon nanotubes and graphene;
- Single-atom transistor.* Develop a single-atom transistor;
- Quantum dots.* Design logic around quantum dots;
- Spintronics.* Employ spintronics to represent and process data;
- DNA computing.* Employ DNA computing to represent and process data; and
- Quantum computers.* Employ quantum computers to represent and process data.

Although the last three might produce benefits only in specialized cases (such as certain massively parallel search problems), some of them are so widespread that special processors may be economically viable.

An IEEE group called "Rebooting Computing" (<http://rebootingcomputing.ieee.org/>) is examining how to continue technology scaling in a post-Moore's Law era.²⁹ John Shalf of the National Energy Research Scientific Computing Center and Robert Leland of Sandia National Laboratory discussed a comprehensive study to form a taxonomy of possible CMOS-successor technologies,²⁹ identifying five possible categories in which successors might be found:

- Architectures and software advances.* Energy management, new kinds of circuits, system on a chip, neuronal chips,¹⁹ specialized chips, dark silicon, and near-threshold voltage operation;
- 3D integration and packaging.* Multiple tiered stacked chip, metal layers, and other types of active layers;
- Resistance reduction.* Superconductors and crystalline metals;
- Millivolt switches (better transistors).* Tunnel field-effect transistors, heterogeneous semiconductors, carbon nanotubes, graphene, and piezoelectric transistors; and
- New logic paradigms.* Spintronics, topological insulators, nanophotonics, biological computing, and chemical computing.

It is noteworthy that many of these directions involve technology jumping, a phenomenon observed in the Kurzweil charts,²⁵ as in Figure 3. The search for CMOS successors aims to jump to a new technology and continue the exponential growth from there.

Wu et al. are confident these lines of development will produce exponential growth advances for another 50 years. Shalf and Leland are more cautious but still show considerable optimism. Only time will tell, but you can be sure that some very good people are working each of these angles.

a balanced H-tree on a new layer of the chip,²⁷ but layering is controversial because putting other circuits (such as memory) on the new layer might benefit computing capacity more.

Due to these practical difficulties with clock-signal distribution, many designers have looked to asynchronous circuits. Two subsystems can exchange data by following a ready-acknowledge protocol. The sender signals that it has some data to send by setting a "ready" line to 1. On receipt of the signal, the receiver acknowledges by setting an "acknowledge" line to 1. When the sender sees the acknowledge, it deposits the data in a buffer and signals completion by returning "ready" to 0. Finally, the receiver takes the data from the buffer and sets "acknowledge" to 0. One unit of data can be transmitted at each cycle of this protocol.

Circuit designers have studied asynchronous signaling since the 1960s and developed reliable asynchronous circuits. Because these circuits are somewhat slower than clocked circuits, ready-acknowledge signaling is used only when there is no alternative (such as a CPU interacting with an I/O device).

Modern chips sometimes use asynchronous signaling (judiciously) to overcome clock distribution and skew problems. The chip is divided into modules, each with its own clock. Clocked circuits are used inside a module with asynchronous signaling between modules. For example, the module that displays a graphical image can be turned on just when a user wants to display the image and run at the clock speed necessary to render the image well; communication with the display module can be asynchronous.

Designing an all-asynchronous computer has been a holy grail among circuit designers for years. Intel is said to have demonstrated an asynchronous circuit for the fetch-execute control of a CPU but has not yet succeeded at creating an arithmetic-logic unit (ALU). The main research problem is finding a way for the ALU to report when it is done with an operation, given that the time of the operation can vary significantly depending on the inputs. Computer graphics pioneer Ivan Sutherland has long advocated for all-asynchronous circuits and demonstrated asynchronous pipeline chips.³⁰

Perhaps the most complete design of a computer that has no clocks is the dataflow architecture proposed by computer scientist Jack Dennis.⁹ It did not inspire sufficient commercial interest because its circuits ran slower than conventional clocked circuits; the machine got its throughput from massive data parallelism, which was not common at the time. Because the number of data-intensive applications continues to grow, the massively parallel dataflow architecture may yet find acceptance.

Potential technology directions. Many engineers have been studying how to enable the continued growth of Moore's Law, given that the existing CMOS technology cannot be pushed much further; for more on the intensive research in this area, see the sidebar "Technology Jumping in Pursuit of Moore's Law."

Exponential Growth at the System Level

Users of computation are hungry for performance, measured as calculations per second or (more recently) calculations per watt-hour. But it does little good to embed faster chips in systems that are limited by other bottlenecks (such as communication bandwidth and cooling systems).

Bottlenecks are the main barrier to performance in systems. Engineers spend a lot of time identifying bottlenecks and speeding them up. Each generation of system improvement is more challenging because engineers must search multiple new technologies to resolve all bottlenecks.

Colwell⁵ discussed bottlenecks generated by "neighboring technologies," or technologies from other fields on which microchips depend. Wu et al.³³ gave a nice example with ubiquitous modern analog-digital converters (ADC). ADCs sample a continuous input signal at twice its highest frequency, producing a series of digital snapshots; according to the Nyquist sampling theorem, no information is lost at this sampling rate because the continuous signal can be regenerated from the samples. As logic circuits get faster, the ADC sampling rate is itself eventually a bottleneck. Engineers are searching for new sampling methods with higher rates.

The memory system is another potential bottleneck. Caches are a critical driver

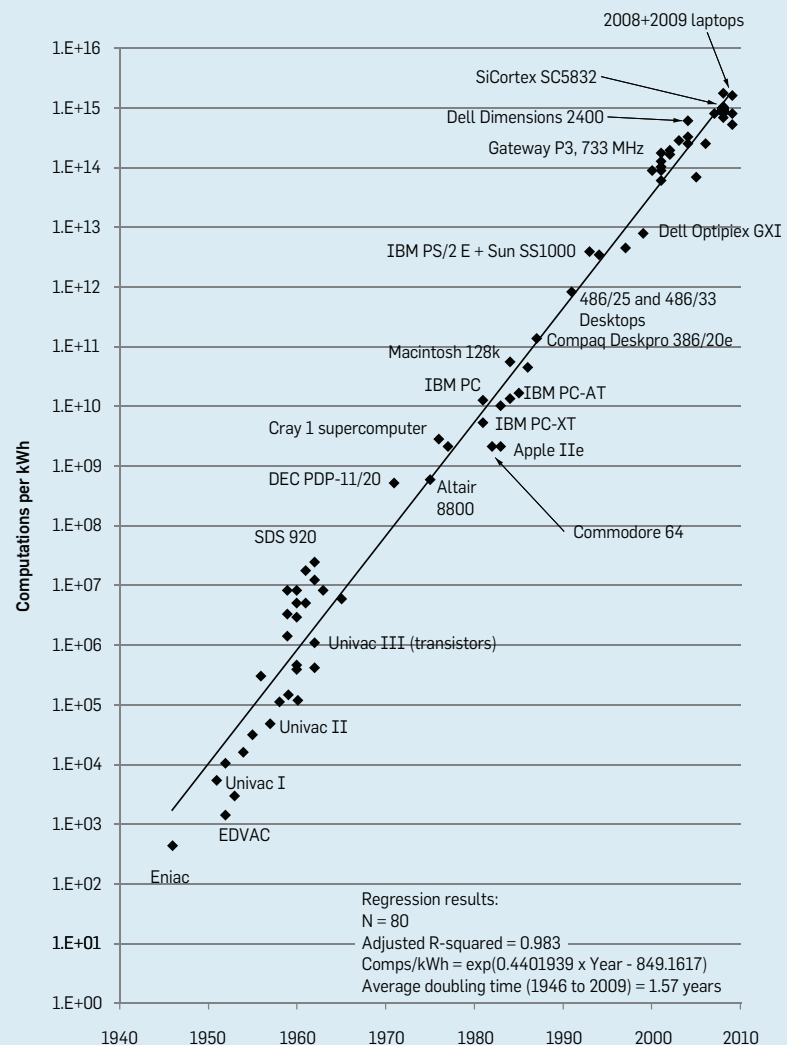
of performance; modern chips rely extensively on caches to position needed data near the processor, and poorly positioned data slows the cache and, in turn, the processor. Considerable research effort has gone into measuring locality of workloads and designing the caches for optimal performance with those workloads. Cache designers are under constant pressure to produce memory improvements matching CPU improvements.

Despite these challenges, computer engineers have been very successful over the years at producing complete systems with performance that has grown exponentially. Koomey^{22,23} gathered data for a large number of different

computers from 1946 to 2009 and found exponential growth in two measures: computation speeds per computer and computations per kilowatt-hour (kwh); see Figure 8 for his graph of computations per kwh. The doubling times were about the same—1.57 years—in both graphs. The improvements relative to energy consumption, summarized as Koomey's Law,¹⁰ have assumed great importance in an energy-constrained world, from large data centers with fixed power draw to mobile devices with fixed battery life. This trend could continue for at least another several decades.

Even when systems are designed so no bottlenecks stand in the way of

Figure 8. Koomey's Law graph illustrates the continuing success of designing systems that produce more computation for the same power consumption. Careful power management over the past decade has enabled the explosion of mobile devices that depend critically on technologies that minimize power use.



performance, it is possible that the workloads presented to those systems are not sufficient to use all the available computing power. In 1967, Gene Amdahl, a mainframe computer designer at IBM, investigated whether it would be better to get faster speed through a faster CPU or through several slower parallel CPUs. Based on his experience designing instruction-lookahead CPUs, he realized that substantial parts of code must be executed sequentially in the given compiled sequence; only some of the instructions could be speeded up by parallel execution. He derived a formula that became known as Amdahl's Law to express the speedup potential from a set of n processors (cores) working on a program. Amdahl's idea—expressing the parallelizable part as a set of parallel instruction streams—was known in his time as “control parallelism.” Suppose the time of the job using 1 stream is $T(1) = a+b$, where a is the time for the serial part, and b is the time for the parallelizable part. The serial fraction is $p = a/(a+b)$, and parallelizable fraction is $1-p$. The time to execute on n streams is $T(n) = a + b/n$ because only the control-parallel portion of the algorithm can benefit from n processors. Amdahl's Law says the speedup is

$$(1) \quad S_A(n) = \frac{T(1)}{T(n)} = \frac{a+b}{a+b/n} = \frac{n}{np + (1-p)} < \frac{1}{p}$$

For example, an application 10% serial would run at most 10 times faster than its single-stream time, even with a large number of parallel processing cores. That is, even if a small part of the overall computation is serial, it is impossible to achieve much multicore speedup with control parallelism.¹⁸

Amdahl's Law would seem to limit the speedup to considerably less than the number of cores, because it assumes parallelism comes from remodeling a serial algorithm into a parallel algorithm. However, Amdahl's Law overlooks parallelism inherent in data. Data-parallel workloads are now common in data-intensive applications (such as MapReduce operations on the Web). In a data-intensive problem, the data space can be partitioned into many small subsets, each of which can be processed by its own thread. The finer the grain of the partition, the larger the number of threads. The same algorithm runs in each thread on the data subset

belonging to that thread. Computer scientist John Gustafson observed that large data-intensive problems could always be partitioned into as many grains as could be supported by cores in the processors when there is sufficient data parallelism.¹⁶ In this case, the computational work completed on one core is $W(1) = a+b$, as outlined earlier. With n cores, it jumps to $W(n) = a + bn$ because each of the n cores is performing the same operation on its thread's (different) data items. Gustafson's Law says the speedup is linear in n

$$(2) \quad S_G(n) = \frac{W(n)}{W(1)} = \frac{a+bn}{a+b} = \frac{a+b}{p+n(1-p)} > n(1-p)$$

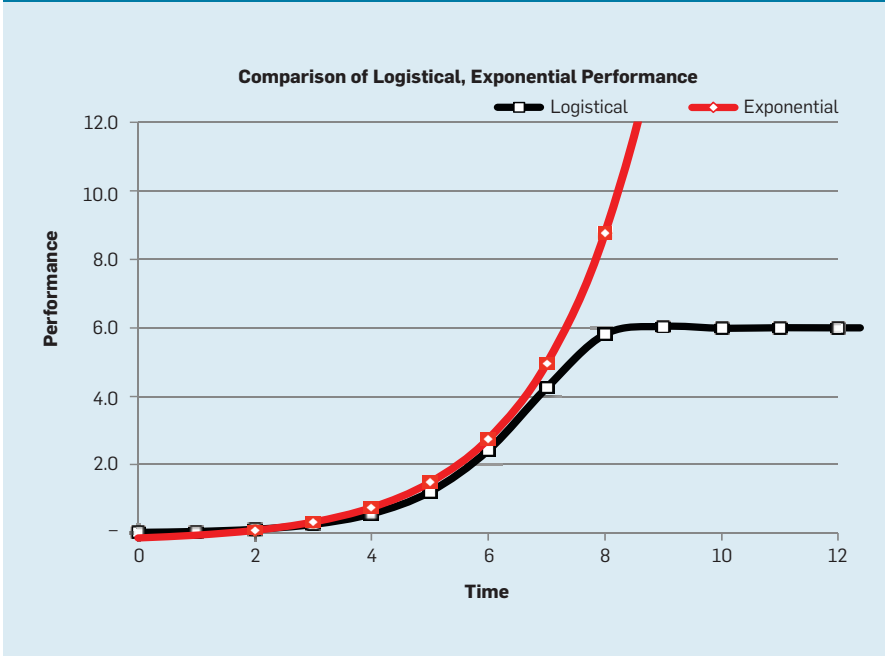
That is, for data-intensive applications with small serial fraction p , adding cores increases the computational work in direct proportion to the number of cores.

Rather than parallelizing the algorithm, data-parallel programming parallelizes the data. Gustafson's Law models data-parallel computing, where the speedup scales with the size of the data, not the number of control-parallel paths specified in the algorithm.

Parallelizing data instead of algorithms was a paradigm shift that began in the 1980s.⁸ Today, it exploits multicore systems even when algorithms cannot be parallelized. It extends to heterogeneous data parallelism for common applications on the Internet (such as querying databases, serving email, and executing graphics-intensive applications, as in games). Peer-to-peer computing is a form of loosely coupled data parallelism, and cloud computing with multicore servers is a form of tightly coupled data parallelism.

Data parallelism and its variants is why multicore systems can continue to double output without increasing clock frequency. At the system level, as long as the applications contain many parallel tasks, there is always work available for the new cores in next-generation systems. This paradigm is especially useful for processing big data now being routinely provided by users of products from companies like Google, Facebook, Twitter, and LinkedIn.

Figure 9. The logistics function—the mathematical model for growth of a population (such as adopters of a technology)—plots as an S-curve (black) over time. Initially, the curve follows an exponential (red) curve, but after an inflection point (here at time 6) it flattens out because of market saturation.



Technology Diffusion at the Community Level

Moore spoke of a second law, also known as Rock's Law, which is less well known than Moore's own component-doubling law. The second law says that the cost of the fabrication facility for new chips doubles approximately every four years. This is due to the greater precision and ever-smaller size of lithography. An important implication is that the market for a new generation of chips at the same price must be at least double the market for the current generation, just to pay for the new fabrication facility. That is, Moore and Rock recognized that the markets had to expand exponentially to support the continuation of the basic Moore's Law. Without exponential expansion of adoption at the community level, Moore's Law at the chip level would be unsustainable.

Many business strategists believe in the S-curve model whereby the number of people using a technology initially grows exponentially to an inflection point and then flattens out (see Figure 9). The flattening out is caused by the saturation of the market—no more new adopters. Businesses try to time their entry into new technologies whose new S-curves are in their exponential growth stage when the older technology starts to flatten, or “tech-

nology jumping.” Businesses ride a series of S-waves and experience continuous exponential growth as they hop from one wave to the next.

Technology jumping is an integral, recurrent theme in computing. We noted earlier that Kurzweil explained exponential growth in the power of information technology by five massive switches to new technology that made older ones obsolete.²⁵ He assumed that the process of technology jumping will continue well into the 21st century. Steve Jobs of Apple spoke frequently about his strategy of timing his jump to the next technology with the inflection point of the S-curve for the current technology. Andy Grove of Intel took the emergence of a new technology that did a job 10 times better than the current technology as a sign of an inflection point and built his company's strategy around well-timed jumps.¹⁵

Whether or not a new technology is adopted depends on whether people use it instead of something else to accomplish something they care about.¹³ Innovators play an important role in this process by making products and services that influence community members to commit to adopt the new technology into their practice.

A simple argument shows why initial growth of adoption is exponential. Sup-

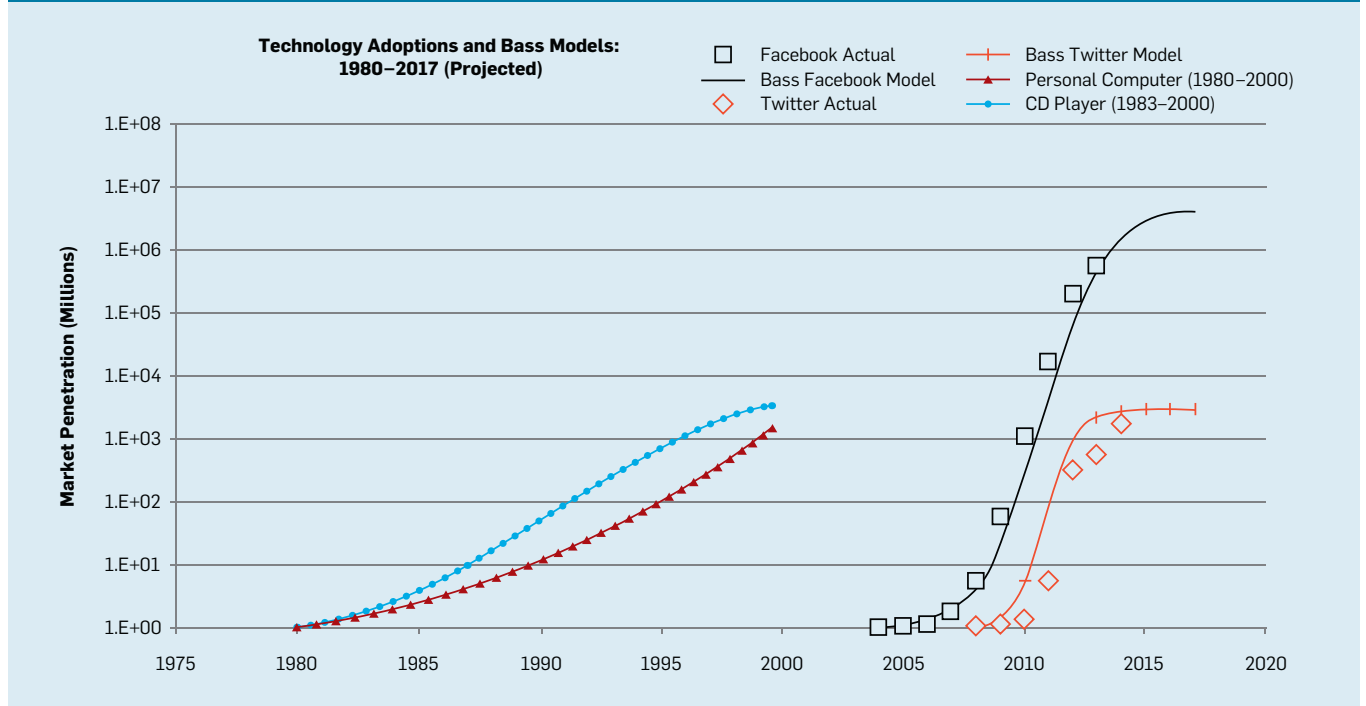
pose $n(t)$ is the number of members of a population who have adopted a technology as of time t . Each adopter demonstrates the value of the new technology to the rest of the population. Let a be the rate at which one adopter influences a non-adopter to adopt. In a small time interval h , the probability that a new adopter comes on board is ah . At time $t+h$, there are thus $ahn(t)$ new adopters, giving $n(t+h)-n(t) = ahn(t)$. Rearranging this equation and letting h go to 0 yields the differential equation

$$(3) \quad \frac{d}{dt} n(t) = an(t)$$

The solution to this equation is $n(t) = e^{at}$. The size of the adopting population increases exponentially and the mean time between adoptions is $1/a$. We can conclude from this simple equation that exponential growth happens when the rate of change also increases with current state.

This model is also too simple because it does not account for diminishing returns due to market saturation or technology reaching a limit where growth stops, when, for example, everyone in the population has adopted the technology. We can extend the model to account for the diminishing population of non-adopters. As before, let $n(t)$ denote the number of members

Figure 10. Four different technology-adoption histories illustrate how versatile the Bass model is for accurately representing and forecasting technology adoptions. Data gleaned from Gentry,¹⁴ Jones,²¹ and Kumar.²⁴



in a population of size N who have adopted a technology. The change in the number of adopters from time t to time $t+h$ would then be proportional to two quantities: the number who have already adopted, as in equation 3, and the fraction who have not yet adopted. This gives the differential equation

$$(4) \quad \frac{d}{dt}n(t) = an(t)\left(\frac{N-n(t)}{N}\right)$$

which is called the “logistics equation” in the literature; its solution is

$$(5) \quad n(t) = \frac{N}{1 + (N-1)e^{-at}}$$

This function grows exponentially to its inflection point. Initially, when $t=0$, there is only one adopter, $n(0) = 1$, and after a long period of time there are N adopters.

Reality is more complicated because individuals have their own adoption time constants. Sociologist Everett Rogers discovered in 1962 that individuals fall into five groups according to the time they take to commit to an adoption—innovator, early adopter, early majority, late majority, and laggard.²⁸ The histogram of adoption times follows a Bell curve. The five categories correspond to five zones of standard deviations. For example, innovators are 2.5% of the population, with adoption times at least two standard deviations below the mean; early adopters are 13.5% and are one to two standard deviations below the mean; early majority are 34% and zero to one standard deviations below the mean. In 1969, professor of marketing Frank Bass modified the Rogers diffusion model by quantifying the impact of early-adopter and word-of-mouth (all other) followers, inserting parameters p (early adoption rate), q (word-of-mouth follower rate), and N into the simple logistics curve.^{3,4} Ashish Kumar and others have since validated Bass’s extensions for real products by finding parameters p , q , and N for a number of technology products.^{14,24} Setting $a=p+q$ and $r=q/p$, the Bass model then gives

$$(6) \quad n(t) = \frac{N(1 - e^{-at})}{1 + re^{-at}}$$

Bass’s equation is still a logistics model that more accurately forecasts sales and also reaches an inflection point where the exponential growth

Data parallelism and its variants is why multicore systems can continue to double output without increasing clock frequency.

begins to slow as diminishing returns set in. Figure 10 depicts four different technology adoptions since 1980. Bass’s model can be solved to find the inflection-point value of at , helping fit the model to data.

No matter what, the initial growth up to the inflection point is exponential. Historically, information technologists have jumped from one technology to another when an incumbent technology nears its inflection point. After each jump, they are on a new curve growing exponentially toward a new inflection point. Moore’s Law, and others like it, are intrinsically exponential because their rate of change is proportional to their current state.

Conclusion

The original 1965 Moore’s Law was an empirical observation that component density on a computer chip doubled every two years. Similar doubling rates have been observed in chip speeds, computer speeds, and computations per unit of energy. However, the durability of these technology forecasts suggests deeper phenomena.

We have argued that exponential growth would not have succeeded without sustained exponential growth at three levels of the computing ecosystem—chip, system, and adopting community. Growth (progress) feeds on itself up to the inflection point. Diminishing returns then set in, signaling the need to jump to another technology, system design, or class of application or community.

At the chip level, there are strong economic motivations for chip companies and their engineers to feed on previous improvements, building faster chips that grow exponentially up to the inflection point. The first significant technology path for exponential chip growth appeared in the form of Dennard scaling, which showed how to reduce component dimension without increasing power density. Dennard scaling reached an inflection point in the 1990s due to heat-dissipation problems that limited clock speed to approximately 3.5Ghz. Engineers responded with a technology jump to multicore chips, which gave speedup through parallelism. This jump has been enormously effective. Cloud platforms and supercomput-

ers achieve high computation rates through massive parallelism among many chips.

A major technology barrier to chip growth has been the distribution of clock signals to all components on a chip. The theoretically most efficient distribution mechanism is the space-filling fractal H-tree that reaches diminishing returns when the tree itself starts to consume most of the physical space on the chip. Engineers are experimenting with hybrids that feature subsystems (such as cores) with their own clocks interacting via asynchronous signaling. Some engineers have been exploring the design of all-asynchronous circuits (no clocks), but these systems cannot yet compete in speed with clocked systems.

Engineers have been systematically examining the barriers that prevent the continuation of Moore's Law for CMOS technologies. As alternatives mature, it will be feasible to jump to the new technologies and continue the exponential growth. Although there is controversy about how successful some of the alternatives may be, there is considerable optimism that some will work out and exponential growth can continue with new base technologies.

When chips and other components are assembled into complete computer systems, engineers have located and relieved technology bottlenecks that would prevent the systems from scaling up in speed as fast as their component chips scale. Koomey's Laws document exponential growth of computations per computer and per unit of energy from 1946 to 2009. Koomey's Law for computations per unit of energy is especially important throughout an energy-constrained industry. Additionally, the technology jump from algorithm parallelism to data parallelism further assures us we can grow systems performance as long as the workloads have sufficient parallelism, which has turned out to be the case for cloud and supercomputing systems.

Finally, we demonstrated that simple assumptions about adoption process lead to the S-curve model in which adoptions grow exponentially until an inflection point and then slow down because of market saturation. Business leaders use the S-curve model to guide them in jumping to new technologies when the older ones start to encounter their limits. Exponential growth in

sales through ever-expanding applications and communities provides the financial stimulus to advance the chip- and system-level technologies (Rock's Law). It is a complete cycle.

These analyses show that the conditions exist at all three levels of the computing ecosystem to sustain exponential growth. They support the optimism of many engineers that many additional years of exponential growth are likely. Moore's Law was sustained for five decades. Exponential growth is likely to be sustained for many more.

Acknowledgments

We are grateful to Douglas Fouts and Ted Huffmire, both of the Naval Postgraduate School, Monterey, CA, for conversations and insights as we worked on this article. C

References

- Anderson, R.G. *How Well Do Wages Follow Productivity Growth?* Federal Reserve Bank of St. Louis Economic Synopses, St. Louis, MO, 2007; <https://research.stlouisfed.org/publications/es/07/ES0707.pdf>
- Asanovic, K., Bodik, R., Demmel, J., Keaveny, T., Keutzer, K., Kubitowicz, J., Morgan, N., Patterson, D., Sen, K., Wawrzyniec, J., Wessel, D., and Yelick, K. A view of the parallel computing landscape. *Commun. ACM* 52, 10 (Oct. 2009), 56–67.
- Bass, F. A new product growth model for consumer durables. *Management Science* 15, 5 (Jan. 1969), 215–227.
- Bass, F., Krishnan, T., and Jain, D. Why the Bass model fits without decision variables. *Marketing Science* 13, 3 (Summer 1994), 203–223.
- Colwell, R. The chip design game at the end of Moore's Law. In *Proceedings of the IEEE/ACM Symposium on High-Performance Chips (Hot Chips)* (Palo Alto, CA, Aug. 25–27). ACM Press, New York, 2013; http://www.hotchips.org/wp-content/uploads/hc_archives/hc25/Hc25.15-keynote1-Chipdesign-epub/Hc25.26.190-Keynote1-ChipDesignGame-Colwell-DARPA.pdf
- Dennard, R., Gaensslen, F., Yu, H.-N., Rideout, V.L., Bassous, E., and LeBlanc, A. Design of ion-implanted mosfets with very small physical dimensions. *IEEE Journal of Solid State Circuits* (1974), 256–268.
- Denning, P. The choice uncertainty principle. *Commun. ACM* 50, 11 (Nov. 2007), 9–14.
- Denning, P. and Tichy, W. Highly parallel computation. *Science* 250 (Nov. 1990), 1217–1222.
- Dennis, J. and Misunas, D. A preliminary architecture for a basic data-flow processor. In *Proceedings of the Second Annual Symposium on Computer Architecture (ISCA)* (Houston, TX, Jan. 20–22). ACM Press, New York, 1975, 126–132.
- Economist. A deeper law than Moore's? *Economist* blog (Oct. 10, 2011); <http://www.economist.com/blogs/dailychart/2011/10/computing-power>
- Esmailzadeh, H., Blum, E., St. Amant, R., Sankaralingam, K., and Burger, D. Dark silicon and the end of multicore scaling. In *Proceedings of the 38th International Symposium on Computer Architecture (ISCA)* (San Jose, CA, June 4–8). ACM Press, New York, 2011, 365–376.
- Feynman, R. *The Pleasure of Finding Things Out: The Best Short Works of Richard P. Feynman*. Penguin Books, New York, 2001.
- Flores, F. and Denning, P. Emergent innovation. Interview by Peter J. Denning. *Commun. ACM* 58, 6 (June 2015), 28–31.
- Gentry, L. and Calantone, R. *Forecasting Consumer Adoption of Technological Innovation: Choosing the Appropriate Diffusion Models for New Products and Services Before Launch*. Faculty Research & Creative Works, Paper 662. Missouri University of Science and Technology, Rolla, MO, 2007; http://scholarsmine.mst.edu/faculty_work/662

- Grove, A. *Only the Paranoid Survive*. Doubleday, New York, 1996.
- Gustafson, J. Re-evaluating Amdahl's Law. *Commun. ACM* 31, 5 (May 1988), 522–533.
- Henderson, B. The Experience Curve—Reviewed (Part II). BCG Perspectives (Jan. 1973); https://www.bcgperspectives.com/content/classics/corporate_finance_corporate_strategy_portfolio_management_the_experience_curve_reviewed_history/
- Hill, M. Amdahl's Law in the multicore era. *IEEE Computer* 41, 7 (July 2008), 33–38.
- IBM. Announcement of SyNAPSE Chip: New IBM SyNAPSE Chip Could Open Era of Vast Neural Networks. IBM, San Jose, CA, Aug. 7, 2014; <http://www-03.ibm.com/press/us/en/pressrelease/44529.wss>
- IEEE Spectrum. Special Report on 50 Years of Moore's Law. *IEEE Spectrum* (Apr. 2015); <http://spectrum.ieee.org/static/special-report-50-years-of-moores-law>
- Jones, K. Growth of Social Media v2.0. *Search Engine Journal Blog*, Nov. 15, 2013; <https://www.searchenginejournal.com/growth-social-media-2-0-infographic/77055/>
- Koomey, J., Berard, S., Sanchez, M., and Wong, H. Implications of historical trends in the electrical efficiency of computing. *IEEE Annals of the History of Computing* 33, 3 (July–Sept. 2011) 46–54.
- Koomey, J. More on efficiency trends in computing, from my forthcoming book, blog, including Koomey Law graph and pointer to creative commons attribution-noncommercial-noderivative license, Dec. 11, 2011; <http://www.koomey.com/post/14466436072>
- Kumar, A., Baisya, R.J., Shankar, R., and Momaya, K. Diffusion of mobile communications: Application of Bass Diffusion Model to BRIC countries. *Journal of Scientific & Industrial Research* 66 (Apr. 2007), 312–316.
- Kurzweil, R. *The Age of Spiritual Machines*. Penguin Books, New York, 1999.
- Moore, G. Cramming more components onto integrated circuits. *Electronics* 38, 8 (Apr. 1965), 114–117.
- Panth, S., Samadi, K., Du, Y., and Lim, S.K. Design and CAD methodologies for low-power gate-level monolithic 3D ICs. In *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design* (La Jolla, CA, Aug. 11–13). ACM Press, New York, 2014, 171–176; <http://dx.doi.org/10.1145/2627369.2627642>
- Rogers, E. *Diffusion of Innovations, Fifth Edition*. Free Press, New York, 2003.
- Shalf, J. and Leland, R. Computing beyond Moore's Law. *IEEE Computer* 48, 12 (Dec. 2015), 14–23.
- Sutherland, I. The tyranny of the clock. *Commun. ACM* 55, 10 (Oct. 2012), 35–36.
- Thackery, A., Brock, D., and Jones, R. *Moore's Law: The Life of Gordon Moore, Silicon Valley's Quiet Revolutionary*. Basic Books, New York, 2015.
- Ubiquity. Ubiquity Symposium on the Technological Singularity, 2014; <http://ubiquity.acm.org/symposia.cfm>
- Wu, J., Shen, Y.-L., Reinhardt, K., Szu, H., and Dong, B. A nanotechnology enhancement to Moore's Law. *Applied Computational Intelligence and Soft Computing* 2013, Article ID 426962; <http://dx.doi.org/10.1155/2013/426962>

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, Editor of *ACM Ubiquity*, and a past-president of ACM. The views expressed here are not necessarily those of his employer or of the U.S. federal government.

Ted G. Lewis (tedglewis@redshift.com) is an author and consultant with more than 30 books on computing and hi-tech business, retired professor of computer science, most recently at the Naval Postgraduate School, Monterey, CA, Fortune 500 executive, and co-founder of the Center for Homeland Defense and Security at the Naval Postgraduate School, Monterey, CA.

© 2017 ACM 0001-0782/17/01 \$15.00



Watch the authors discuss their work in this exclusive *Communications* video. <http://cacm.acm.org/videos/exponential-laws-of-computing-growth>

DOI:10.1145/2950044

Central IT needs to guide functional areas and departments toward effective operational and procurement practices.

BY CECIL ENG HUANG CHUA AND VEDA C. STOREY

Bottom-Up Enterprise Information Systems: Rethinking the Roles of Central IT Departments

CROSS-FUNCTIONAL, ENTERPRISE, AND infrastructure systems integrate information across an organization for use by multiple stakeholders.^{14,22} Development or configuration of these large, multimillion-dollar projects is normally overseen by top management in conjunction with the central information technology (IT) department.²¹ There is, however, an increasing number of such systems arising from functional areas or departments in which users work, with central

IT and top management being challenged to implement controls on these systems only after they go live.¹³

Consider the following examples based on interviews conducted in recent years. Half were with chief information officers (CIOs). The other half were with their end users. Our aim was to understand the factors leading to bottom-up enterprise information systems. The first example is from a mid-size hospital in Auckland, New Zealand, where doctors had contracted with a paging service. In the second, university faculty adopted the Moodle learning-management system instead of their locally supported system.

CIO: “This guy ... who built this thing and hosted it out of a literally a garage somewhere in London, and they said, ‘Well, we’d like to trial this.’ And he said ‘fine.’ So they set up a trial using that hosted environment ... they actually implemented it into something like 12 or 16 wards. It was just so successful that after the first month it grew like wildfire.”

Another CIO: “Blackboard was the official system ... People said, ‘We want to bring the Moodle environment test mode, and it will only be in test mode, and we will see how that goes... so they started using lectures on them, so it became a production environment. But it was under their control sitting somewhere on some server, and they’ve got an organization functionally dependent on shadow IT.’”

Bottom-up enterprise information systems are not sanctioned by central IT, the official group within an

» key insights

- **Central IT must deal with the emerging reality of bottom-up enterprise information systems.**
- **Bottom-up enterprise information systems are developed by employees when alternatives are not readily available.**
- **Central IT and functional areas must work together to understand and maintain effective bottom-up enterprise information systems that support overall organizational strategy.**



organization responsible for IT. Nevertheless, they are implemented on an organizationwide or, minimally, a cross-functional basis. They exhibit three key characteristics. First, they are either enterprisewide, or function spanning, so the technology is used across multiple departments. Second, IT implementations are developed, configured, or procured by end users who do not necessarily have sophisticated technology skills. Third, imple-

mentation decisions are made when top management and central IT exert little formal control and have no official governance over the project.

In this article, we investigate the emergent phenomenon of such bottom-up enterprise information systems to identify causal factors and offer recommendations for how to best manage it. Insights are derived from three sources. First and foremost, the paired CIO and end-user interviews

provide concrete examples of how bottom-up enterprise information systems are emerging and being managed from which some general inferences can be made. Second, our interview data is complemented by the practitioner literature from which we derive a framework for understanding the context of bottom-up enterprise information systems. Finally, the theoretical literature, specifically on distributed leadership, provides insights into how

the phenomenon should be managed. The result is a set of recommendations for rethinking the role of central IT to effectively manage bottom-up enterprise information systems.

Drivers of Bottom-Up Systems

The two types of main drivers of bottom-up enterprise information systems are external and internal:

External. The primary reason bottom-up enterprise information systems are emerging today is changes in technology, especially since the mid-2000s, including:

Simpler/“standardized” application distribution models. Historically, enterprise information systems have been configured and installed by specialized vendors. Such applications are now available in downloadable form. For example, ADempiere (<http://adempiere.org>) is an open source enterprise resource planning system; Moodle is an open source learning-management system. Cloud versions of enterprise software are available; for example, SAP Business ByDesign (<http://go.sap.com/product/enterprise-management/business-bydesign.html>) is a cloud-based version of SAP.

Changing IT application cost structure. Historically, new IT applications were expensive capital costs, requiring central IT to secure large budgets in advance.⁶ In contrast, many modern software packages are paid for as ongoing, operational costs; for example, SAP Business ByDesign is accessed through a monthly subscription. Costs are amortized over a longer period,

so functional areas reallocate money from routine budgets to acquire new applications, thereby bypassing financial checks when procuring software.

CIO: “They can reprioritize the definition of what’s an operational expenditure versus a capital expenditure, and, if they have some ‘underspend,’ they can get a developer in and pay just an hourly rate and get some developer time.”

Accessible third-party technical support. One factor impeding the ability of functional areas to procure IT products and support independently has been their lack of IT expertise.¹² However, contemporary applications (such as Microsoft’s Active Directory and TeamViewer) allow monitoring and control of remote computational devices. Functional areas can outsource some IT to a vendor, thereby bypassing both central IT and top management.

User: “... at the moment it’s hosted externally ... it’s an Internet-based system, which means that our staff can access it anywhere, it means ... we can make changes quickly ... Stuff that doesn’t happen through our centralized IT group. Well, it happens, but over a very long period of time.”

Standardization of interfaces. Many enterprise solutions are Web-based, so they work well on most devices. It is thus possible for users to bypass the physical organizational IT infrastructure. Even when organizations block bottom-up enterprise information systems on internal networks, users can still employ personal devices for access.

Together, these external drivers provide users access to complex, multi-function systems that were previously inaccessible.

Internal drivers. There are two main internal organizational drivers:

Inherent limitations of centrally managed governance. Centrally managed governance systems usually favor large, highly visible projects with quantifiable bottom-line benefits. Smaller, possibly exploratory, projects may be neglected. Such projects may have intangible, difficult-to-quantify benefits. The following quote from a CIO illustrates a bias inherent in many centrally managed IT procurement processes.

CIO: “[Person] said we need more e-learning. I said, absolutely, but fat chance that it’s going to make the priority list based on what we are doing. And he said, look I found this online company; it’s an open source product. All I need is \$20,000 to get started.”

This e-learning case highlights little “objective” basis for top management to support such an endeavor. Convincing top management to take notice may indeed be impossible. However, when projects are successful, they can spread quickly, as in the hospital paging service.

Organizational processes. Some organizational processes are ineffective or dysfunctional, making them difficult or costly to work around;²⁴ alternatively, processes may change frequently in today’s dynamic business environment. Functional departments may thus implement bottom-up enterprise information systems because they do not perceive feasible alternatives as being available.

In one Moodle implementation, a particular university’s rules prevented users from enrolling students who attended “short” courses (such as adult education and executive MBA). These students did not go through the complete university student-enrollment process and were not issued university student-identification cards and numbers, a prerequisite for using the existing learning-management system. Users employed Moodle to enroll and manage these students. Here, the central IT system did not cater to a significant (and highly profitable) subsegment of the university’s customer

Changes caused by a bottom-up approach to central IT business processes.

ITIL Process Group	Changes
Service Strategy	Central IT must dedicate more resources to all levels of service strategy: lowest levels focus on functional areas; mid-levels focus on planning and projection; and highest levels focus on negotiation and strategic coordination.
Service Design	Central IT becomes a repository of organizational knowledge and requirements, helping align functional areas’ IT projects with the overall enterprise vision and architecture.
Service Transition	Central IT reminds functional areas to incorporate service transitions into their plans.
Service Operations	Central IT maintains enterprise infrastructure, coordinates across functional areas to ensure operational continuity and quality, and consults with functional areas to facilitate service operations.
Continual Service Improvement	Central IT reminds functional areas to perform CSI, coordinates cross-functional CSI, and advises functional areas on CSI best practices.

base. Functional areas serving that base independently supported them through their own local initiatives.

IS researchers have argued enterprise information systems should not be built in a top-down manner but instead facilitate bottom-up innovation.⁴ Successful implementation of enterprisewide systems (such as enterprise architecture)⁸ will increasingly require negotiation and dialog rather than the imposition of ideas from central IT.²⁵ This bottom-up approach to building enterprise systems now occurs worldwide, regardless of central IT desires. For corporate management, no longer is it an issue of whether bottom-up enterprise information systems are “right” or counterproductive but rather how to deal with the reality of their existence.

Rethinking Central IT

The relevant challenges involve three questions for IT management: How will bottom-up enterprise information systems change the way IT is managed? What should central IT do about the phenomenon? And what is the right mind-set to manage the IT landscape under such change?

Central IT’s role. To address the first two, we adopt concepts from the Information Technology Infrastructure Library (ITIL, <http://www.itlibrary.org/>),^{2,9,15,19,23} a “best practice” industry framework providing an overview of the responsibilities of an IT department. It is well accepted by the IT industry¹⁷ that central IT departments that follow ITIL generally perform better than those that do not.^{10,17} Consistent with the strategic information systems literature (such as in Zoet et al.²⁵), ITIL views the IT function as concerned with services, categorizing IT functions into five groups. The accompanying table summarizes how bottom-up enterprise information systems will change central IT’s practice with respect to these five process groups.

Service strategy. Service strategy is concerned with articulating an IT strategy and aligning IT processes with the overall organizational vision and direction.² Bottom-up enterprise information systems create greater unknowns and uncertainty for IT service strategy; for example, part of the



Central IT needs to transform itself from “geeks on the other floor” to “our IT support next door.”



service strategy is projecting resource needs and ensuring an organization has sufficient capacity to run its future IT. When creating bottom-up enterprise information systems, functional areas may provide their own hardware, databases, and other critical infrastructure. Alternatively, they may employ software-as-a-service in which software is licensed on a subscription basis. Regardless, bottom-up enterprise information systems still consume central IT resources; for example, network load may be increased. If central IT is not aware of bottom-up enterprise information systems, it simply cannot plan for it.

Bottom-up enterprise information systems make human resource planning difficult because service-desk response becomes more complex. The support responsibility for bottom-up enterprise information systems can never wholly lie with functional areas because the infrastructure provided by central IT affects the performance of these systems; for example, an operating system or browser upgrade could be incompatible with a bottom-up enterprise information system, rendering it non-functional. Solving this human-resource-planning problem requires central IT budget for and dedicate people and money to it.

Finally, bottom-up enterprise information systems may not be aligned with an organization’s strategy. The Moodle example violates a strategy of providing students with a consistent experience across all courses. Some faculty members employ Blackboard (<http://blackboard.com/>) and others Moodle. Dealing with such conflicts is challenging for central IT due to difficulties leveraging organizational power to convince functional areas to cooperate. In the Moodle example, faculty members use it knowing Blackboard is the officially sanctioned system. Discussion, understanding users’ constraints (such as when students cannot be registered in the official system), and compromise are all necessary for resolving any ensuing tension.⁴

Dealing with bottom-up enterprise information systems thus requires central IT to devote a greater proportion of its workforce at all levels to service strategy. At the lowest level, IT per-

sonnel must engage with functional areas to understand what they are doing. Information is then transmitted to central IT, enabling better planning. At the middle levels, planning and projection require additional effort by IT personnel. At the highest levels, more effort is required to negotiate and coordinate functional areas to ensure the IT landscape across the organization is in alignment with overall enterprise strategy.

Service design. Service design focuses on development of new services for an organization,⁹ including new software, applications, and infrastructure, as well as acquiring vendor IT products. With bottom-up enterprise information systems, central IT's role in service design is diminished. Nevertheless, there remain two critical roles for IT. First, central IT is a repository of organizational-level requirements that functional areas often do not consider;⁸ for example, in one interview we conducted with a government organization in New Zealand, a user described a contract with a vendor to provide citizen services. We asked how the New Zealand Privacy Act was being addressed, given a vendor had access to the data of private citizens. The interviewee declined to answer. Central IT may be aware of privacy, the regulatory environment, and other broader organizational issues, but the functional areas may not be. It then becomes the responsibility of central IT to educate functional areas about these requirements; for example, in one safety-critical manufacturing organization, the functional areas were reminded that improperly designed IT affecting the manufacturing process could cause the organization to lose its operating license. The organization allows other forms of bottom-up enterprise information systems, as in bottom-up financial IT. However, due to legal requirements, user-developed IT in the manufacturing process is strictly forbidden.

User: "We have to adhere stringently to what we call GMP [good manufacturing process] requirements ... We get audited by [organization] and auditors come in and spend time talking to our IS manager about how those systems are managed and maintained and all that sort of stuff."



For bottom-up enterprise information systems to succeed, central IT must be willing to cede power to functional areas and vice versa.



Second, central IT has a responsibility in guiding and structuring the implementation or vendor-selection process. Vendors tend to have an advantage because they negotiate more contracts for their class of service than do clients.²⁰ Central IT negotiates more IT contracts than users, resulting in greater competence in such matters. Also, central IT is generally more aware of the broader issues in implementation; for example, functions might neglect nonfunctional testing (such as load testing). In the case of the hospital using a paging service mentioned earlier, central IT struggled to ensure that the paging service was supported by reliable infrastructure, an issue users had not considered.

In bottom-up enterprise information systems, although central IT's role in service design might appear diminished, that role remains important. Specifically, central IT becomes the repository of shared organizational knowledge (such as policies for dealing with vendors) and organizational-level (as opposed to functional-level) requirements. Central IT must thus engage with functional areas and/or vendors to ensure bottom-up enterprise information systems remain aligned with the overall enterprise vision and architecture.

Service transition. Service transition includes processes associated with integrating a new service design with existing processes;¹⁹ typical service-transition processes are change management and training. Because functional areas are more entrenched in their own operations, they tend to be competent in service transition. However, they also tend to underestimate the cost and effort required; for example, one of the first ways central IT learns about bottom-up enterprise information systems is when functional areas request assistance to help with maintenance when the original developer leaves an organization. Here, the functional areas failed to adequately plan for training of maintenance personnel.

Central IT's role in service transition is thus to ensure that service transition is planned for and actually occurs. This role is similar to central IT's role in service delivery, in that central IT brings knowledge from previous

projects (such as recognizing that service transition is important) to the new bottom-up-enterprise-information-system implementation.

Service operation. Service operation focuses on ensuring continuity of service.²³ The service desk, ongoing maintenance, ensuring business continuity in the event of a fault, and disaster recovery are all elements of service operation. For bottom-up enterprise information systems, the role of central IT in service operation is complex. First, bottom-up enterprise information systems exist as just one interconnected part in an organization's IT landscape. For bottom-up enterprise information systems to perform well, they must connect to the corporate network, operating system, and firewall, each of which must function well. Furthermore, bottom-up enterprise information systems and these elements must be compatible. As a result, effective service operations for bottom-up enterprise information systems require coordination between central IT and the functional areas. Lack of coordination can indeed create embarrassing situations for CIOs.

CIO "... the guy in Sydney who ran the brand there saw an opportunity to improve on this customer service and he hired a young programmer to come in and build a system for him. ... Eventually, we found out about it because our CEO went there, and he came back to NZ and said to me, 'This is bloody good. Why don't we have it everywhere?,' and I said, 'What is it?'"

Furthermore, functional areas may be unaware of critical elements of effective service operations practices. In the hospital-paging example, the functional areas contracted for service without considering the implications of network or hardware failure. Central IT eventually helped move the paging service onto a network provider that guaranteed a satisfactory level of availability.

With regard to service operations, central IT has three roles. First, it continues to manage much of the enterprise infrastructure (such as networks, servers, and computers). Second, it coordinates with individual functional areas' service desks to ensure continuity of operations across the enterprise. And, finally, it is the repository of orga-

nizational service-desk best practice and advises the functional areas.

Continual service improvement. Continual service improvement explores how organizational IT service delivery can be improved.¹⁵ Managing continual service improvement becomes more complex for central IT because bottom-up enterprise information systems increase the complexity of the IT landscape. Also, because IT assets are owned by functional areas, these areas often must be given a place in planning for continual service improvement. Consider an attempt to rationalize an organization's IT vendors. In a traditional environment, central IT needs to engage only with central IT personnel to do so, because all vendors are managed by central IT. However, with bottom-up enterprise information systems, vendors contract with functional areas, rather than with central IT.

From the perspective of continual service improvement, central IT's role is thus to remind functional areas of the importance of reviewing their services from an improvement perspective, consult on best practices for improving service, and help coordinate improvements across the enterprise. By implication, in environments where bottom-up enterprise information systems are employed, central IT increasingly becomes a coordinating and advisory body, representing higher-level organizational interests.⁸ Operational concerns (such as design, implementation, maintenance and support, change management, training, and vendor management) devolve to functional areas. Note that central IT continues to have an operational role, albeit a diminished one compared to traditional roles with no bottom-up development. At the strategic level, central IT transforms from a function that declares the appropriate technology platforms to a consultative function.

Central IT Leadership

This coordinating and advisory role requires a change of mind-set to succeed. IT leadership may no longer be the purview of the central IT function; rather, it will be distributed across functional areas. IT consulting firm Gartner (<http://www.gartner.com/technology/home.jsp>) has, since 2013, suggested that marketing within an organization may spend more on IT than central IT would spend.⁷ Successful distributed leadership depends on four critical ingredients: balance of power, role blurring, trust, and trusteeship.^{1,5}

Balance of power. Balance of power is created by results from bottom-up enterprise information systems. Central IT, the functional areas, and the vendor (if one exists) can all be the source of unsatisfactory outcomes. Balance of power means increasingly informal, socially driven mechanisms, or relational governance¹⁸ will be applied to manage IT issues. It is difficult for IT management to exert formal power over others who possess similar degrees of formal power or exercise it on those not in a direct reporting line (such as central IT to the functional areas).

A promising foundation for effective relational governance is strong social capital consisting of three equally important ingredients:^{3,11}

Physical connection. Structural social capital is associated with one's ability to physically connect with others; for example, if central IT is physically located on a floor or building different from users, structural social capital is generally low;

Shared language. Cognitive social capital is associated with shared language. Central IT thus has better social capital with the accounting department if it can use accounting terms correctly. However, cognitive social capital is not simply associated with professional language; for example, central IT personnel who graduate from the same university as the accountants are likely to facilitate cognitive social capital; and

Building bonds. Relational social capital focuses on building bonds through favor exchanges, positive interactions (such as shared parties), and trust-building activities.

Central IT must carefully weigh its social capital in this new landscape; for example, it may not want to isolate itself on a separate floor from the functional areas. Similarly, it may want to consider whether the language it uses or the culture within central IT isolates it from other functional areas and

from senior management. Central IT needs to transform itself from “geeks on the other floor” to “our IT support next door.”

Role blurring occurs when one party can replace others if required.^{1,5} Role blurring requires personnel within central IT to gain an accurate understanding of the nature of the functional areas. They must have sufficient technical capability to understand the vendor. Likewise, IT personnel within the functional areas must expand their IT skills so they can carry out, say, system design. Similarly, vendors must have a degree of understanding of the businesses with which they engage.

An organization having bottom-up enterprise information systems is likely to require IT professionals to acquire a more “T-shaped” skill profile, with broad knowledge of a wide variety of skills and deep knowledge in one area.¹⁶ Role blurring is related to cognitive social capital. Central IT personnel must understand the language of the functional areas, even as they communicate their concerns effectively.

Trust. Trust is the degree to which each party believes the other will act in good faith. This requires central IT and the functional areas to build rapport. Interestingly, the IT personnel who often have the most contact with users work with service support, making them best positioned to initiate the building of rapport. However, many central IT departments often treat service support as strategically unimportant, as evidenced, when, say, they outsource it. Central IT might consider how doing so denies itself the opportunity to better engage with functional areas to build necessary rapport.

Trusteeship. Trusteeship requires each party to have the ability to audit and veto actions of the other parties. For bottom-up enterprise information systems to succeed, central IT must be willing to cede power to functional areas and vice versa. Trusteeship is important due to the greater degree of coordination required to make bottom-up enterprise information systems truly successful. Functional areas and central IT must be able to see into each others’ processes to ensure IT systems are not duplicated and are able to work together. Functional areas in organizations with bottom-up enterprise in-

formation systems must be open and transparent with central IT, ensuring this coordination occurs.

User: “... this committee was [the CIO]’s idea. Let’s understand what [bottom-up enterprise information systems] are out there. Let’s gather all the information. We are not going to hit people on the head. We want them to be open to say, ‘What have you got? We are going to try and help you. We’ve got a better solution or support the solution you’ve got.’”

Conclusion

This article has examined bottom-up enterprise information systems to understand why they are being developed and suggest how they might be managed more effectively. Modern technology (such as open source software, outsourcing, and cloud computing) enable users to bypass central IT to procure or configure their own enterprise information systems. Coupled with inherent organizational limitations, the result is bottom-up enterprise information systems becoming a new reality in many organizations. Managing such systems requires central IT to be a collaborative partner that guides functional areas toward effective IT practices. Central IT’s operational role may diminish as the functional areas assume increased duties. However, when IT roles are distributed across the organization, additional coordination is required by central IT, functional areas, and vendors alike. Practices based on distributed leadership include establishing a balance of power across central IT and the functional areas, role blurring, or having both central IT and functional areas learn one another’s skills and concepts, develop trust, and assume trusteeship to help ensure success in contemporary organizational environments.

Acknowledgment

This research was supported by DesignerTech, the University of Auckland, and the J. Mack Robinson College of Business at Georgia State University.

References

1. Bolden, R. Distributed leadership in organizations: A review of theory and research. *International Journal of Management Reviews* 13, 3 (Sept. 2011), 251–269.

2. Cannon, D. *ITIL Service Strategy*. The Stationery Office, Norwich, U.K., 2011.
3. Chua, C.E.H., Lim, W.K., Soh, C., and Sia, S.K. Enacting clan control in complex IT projects. *MIS Quarterly* 36, 2 (June 2012), 577–600.
4. Ciborra, C. From thinking to tinkering: The grassroots of strategic information systems. *The Information Society* 8, 4 (1992), 297–309.
5. Denis, J.-L., Langley, A., and Sergi, V. Leadership in the plural. *Academy of Management Annals* 6, 1 (June 2012), 211–283.
6. Dos Santos, B.L. Justifying investments in new information technologies. *Journal of Management Information Systems* 7, 4 (Spring 1991), 71–90.
7. Golden, B. As CMOs grab IT budget from CIOs, cloud CapEx and OpEx shift. *CIO Magazine* (Jan. 28, 2013).
8. Greefhorst, D. and Proper, E. *Architecture Principles: The Cornerstones of Enterprise Architecture*. Springer-Verlag, Berlin, Germany, 2011.
9. Hunnebeck, L. *ITIL Service Design*. The Stationery Office, Norwich, U.K., 2011.
10. Iden, J. and Eikebrokk, T.R. Using the ITIL process reference model for realizing IT governance: An empirical investigation. *Information Systems Management* 31, 1 (2014), 37–58.
11. Kirsch, L.J., Ko, D.-G., and Haney, M.H. Investigating the antecedents of team-based clan control: Adding social capital as a predictor. *Organization Science* 21, 2 (Apr. 2010), 469–489.
12. Ko, A.J., Abraham, R., Beckwith, L., Blackwell, A., Burnett, M., Erwig, M., Scaffidi, C., Lawrence, J., Lieberman, H., Myers, B., Rosson, M.B., Rothermel, G., Shaw, M., and Wiedenbeck, S. The state of the art in end-user software engineering. *ACM Computing Surveys* 43, 3 (Apr. 2011), 1–44.
13. Lacey, C. Shadow IT Goes Mainstream. Unisys blog, May 6, 2015; <http://blogs.unisys.com/financialindustryinsights/shadow-it-goes-mainstream/>
14. Liang, H., Saraf, N., Hu, Q., and Xue, Y. Assimilation of enterprise systems: The effect of institutional pressures and the mediating role of top management. *MIS Quarterly* 31, 1 (Mar. 2007), 1–29.
15. Lloyd, V. *ITIL Continual Service Improvement*. The Stationery Office, Norwich, U.K., 2011.
16. Madhavan, R. and Grover, R. From embedded knowledge to embodied knowledge: New product development as knowledge management. *Journal of Marketing* 62, 4 (Oct. 1998), 1–12.
17. Pollard, C. and Cater-Steel, A. Justifications, strategies, and critical success factors in successful ITIL implementations in U.S. and Australian companies: An exploratory study. *Information Systems Management* 26, 2 (Apr. 2009), 164–175.
18. Poppo, L. and Zenger, T.R. Do formal contracts and relational governance function as substitutes or complements? *Strategic Management Journal* 23, 8 (May 2002), 707–725.
19. Rance, S. *ITIL Service Transition*. The Stationery Office, Norwich, U.K., 2011.
20. Saunders, C., Gebelt, M., and Hu, Q. Achieving success in information systems outsourcing. *California Management Review* 39, 2 (Dec. 1997), 63–79.
21. Simonsen, J. Involving top management in IT projects. *Commun. ACM* 50, 8 (Aug. 2007), 53–58.
22. Star, S.L. and Ruhleder, K. Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research* 7, 1 (Mar. 1996), 111–134.
23. Steinberg, R. *ITIL Service Operation*. The Stationery Office, Norwich, U.K., 2011.
24. Wagner, J.A. The organizational double bind: Toward an understanding of rationality and its complement. *Academy of Management Review* 3, 4 (Oct. 1978), 786–795.
25. Zoet, M.M., Heerink, A.W., Lankhorst, M.M., Hoppenbrouwers, S.J.B.A., and Stokkum, W.v. An agile way of working. In *Agile Service Development: Combining Adaptive Methods and Flexible Solutions*. Springer-Verlag, Berlin, Germany, 2012, 111–140.

Cecil Eng Huang Chua (aeh.chua@auckland.ac.nz) is an associate professor in the Business School of the University of Auckland, Auckland, New Zealand.

Veda C. Storey (vstorey@gsu.net) is the Tull Professor of Computer Information Systems in the J. Mack Robinson College of Business and Professor of Computer Science at Georgia State University, Atlanta GA.

Blog Ubiquity

INFORMATION EVERYWHERE



The newest ACM forum.

Contributions
that cover the vast
information-rich world
where computing
is embedded
everywhere.

ACM's *Ubiquity* is
the online magazine
oriented toward the
future of computing
and the people who
are creating it.

We invite you to
participate: leave
comments, vote for
your favorites, or
submit your own
contributions.

Captivating topics.

*Net Neutrality
and the Regulated
Internet*

*The End of Life
As We Know It*

*A Shortage of
Technicians*

*The Fractal
Software
Hypothesis*

*Your Grandfather's
Oldsmobile—NOT!*

*Superscalar
Smart Cities*



Association for
Computing Machinery

Visit us at
<http://ubiquity.acm.org/blog/>

Cell-graph construction methods are best served when physics-driven and data-driven paradigms are joined.

BY BÜLENT YENER

Cell-Graphs: Image-Driven Modeling of Structure-Function Relationship

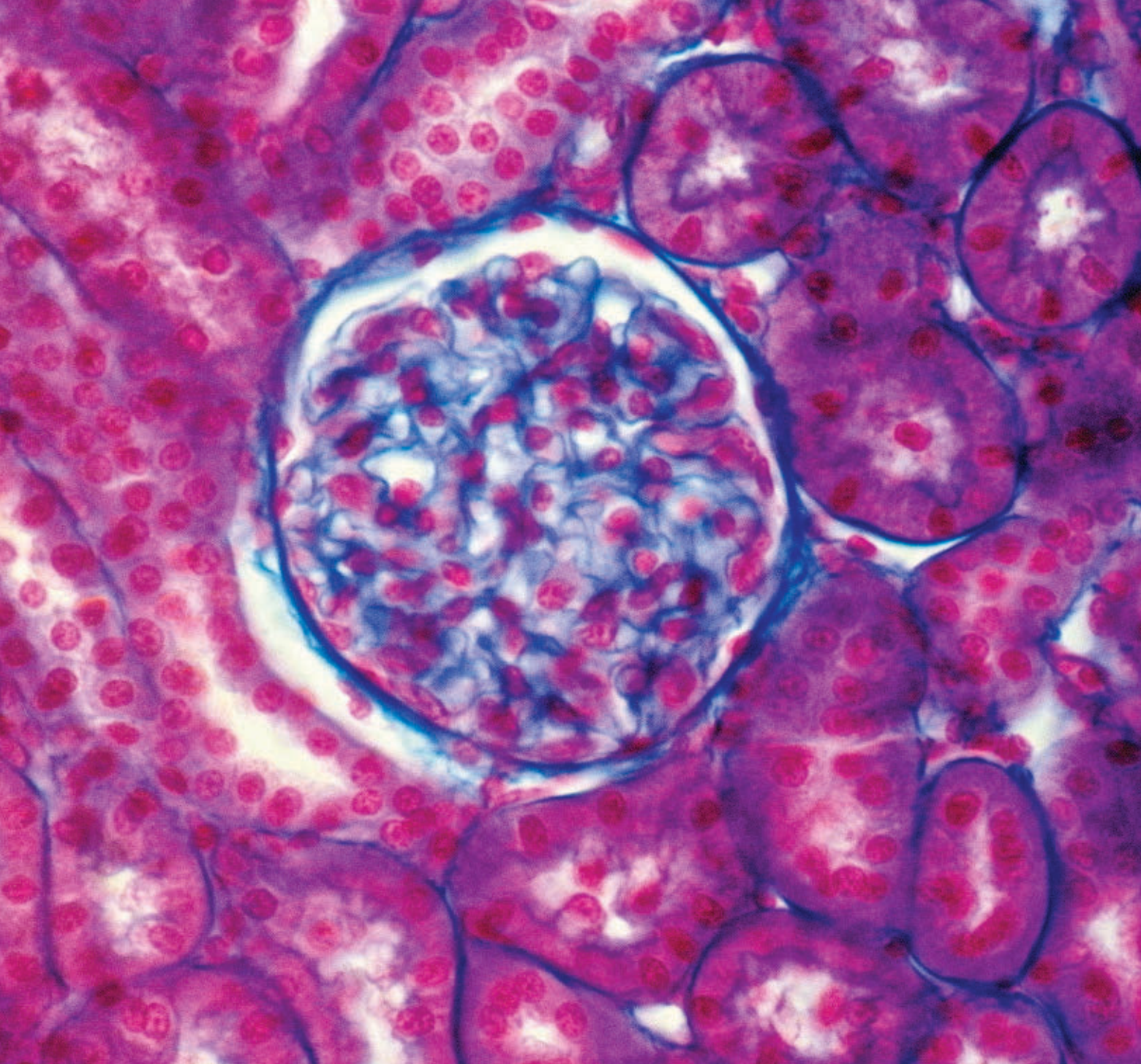
THE STRUCTURE-FUNCTION RELATIONSHIP is fundamental to our understanding of biological systems at all levels, and drives most, if not all, techniques for detecting, diagnosing, and treating a disease. The predominant means of collecting structure/function data in biomedicine is reductionist and has thus led to a proliferation of complex data (for example, gene expression arrays, digital images) that captures only a fraction of the structure/function relationship. Gene sequence and expression data illustrates the structure and activities of individual genes but does not explain how these genes collaborate to control cellular and tissue-scale functions. As a result, despite the

abundance of molecular details known about wound healing, for example, it is virtually impossible to accurately predict the final functional state of a healing wound.³⁶ This illustrates a need to build models that represent the structural organization at the organ, tissue, cellular, and molecular levels. Furthermore, such models must capture relationships between these scales and relate them to the underlying functional state.

Data-driven network/graph analysis is primed to decipher cellular interactions in the intricate relationship between protein-protein interactions, genetic changes, metabolic pathways, and chemical secretions, which comprise cellular events. When extended to the organ level, the key challenge would be to link the local and global structural properties of tissues to the overall morphology and function of a tissue. Only a systems-level understanding of the various cellular processes encompassing multiple biological levels will take into account the multidimensional complexity of these processes. If the principles governing biological organization on a morphological, spectral, local, and global scale can be deduced, the correlation between structural and molecular signaling within the tissue can be understood and applied to inform and accelerate studies of organ development and tissue regeneration.

» key insights

- **Structural and spatial patterns of cell organizations in a tissue are not random but associated with the underlying functional state. Cell-graphs combine techniques from image analysis, graph theory, data mining, and machine learning to identify such patterns to predict underlying functional state. Thus, understanding structure-function relationships can be used to predict malfunctioning when the patterns start changing.**
- **Advances in tissue staining and image processing permit capturing multichannel, multiscale information which in turn can be used by the state of the art machine learning algorithms to model structure-function relationships.**



The *cell-graph* technique^{11,12,20} aims to learn structure-function relationship by modeling structural organization of a tissue/organ sample using graph theory. Its main hypothesis is that cells in a tissue/organ organize to perform a specific function. For example, the spatial distribution and interaction of cells in a salivary gland tissue is different than that of a brain tissue since they perform very different functions. Thus, if one can understand tissue organization then one can successfully predict the corresponding function. The cell-graph technique deploys image processing, feature extraction and selec-

tion, and machine learning algorithms to establish a quantitative relationship between structure and function.

As more sophisticated staining techniques that provide information about different biological scales are deployed (as will be discussed), image-driven modeling with cell-graphs provides a multiscale approach to modeling complex biological systems, as a complementary one to physics-based continuous models and methods (for example, finite element method, fine difference method). While quite successful in various engineering applications, these methods operate under computational scales

that capture the macroscale behavior of a system by smaller (micro) scale constitutive relations. For example, the Car-Parrinello Molecular Dynamics (CPMD) model employs a “microscopic” model to formulate the “constitutive relations” based on a force field between the nuclei, with a “macroscopic” model that uses mechanics for the dynamics of the nuclei.⁷ However, complex biological systems have different scales, including molecular, cellular, tissue, and organ levels, than the computational ones. Furthermore, physics-based techniques are parametric and do not leverage the massive amounts of data available due

to advances in data acquisition, such as high throughput medical imaging techniques, which has been recognized as an important research direction (see <https://datascience.nih.gov/bd2k>).

In this article, I will illustrate various cell-graph construction methods for different applications, explain the graph features used in tissue classification, and suggest how to combine physics-driven and data-driven paradigms toward a multiscale modeling for better prediction. The discussion starts from a simple graph model and progresses toward more sophisticated ones as a function of staining, and imaging techniques. This review includes both static data and time-evolving dynamic data as well.

Image-driven native tissue modeling.

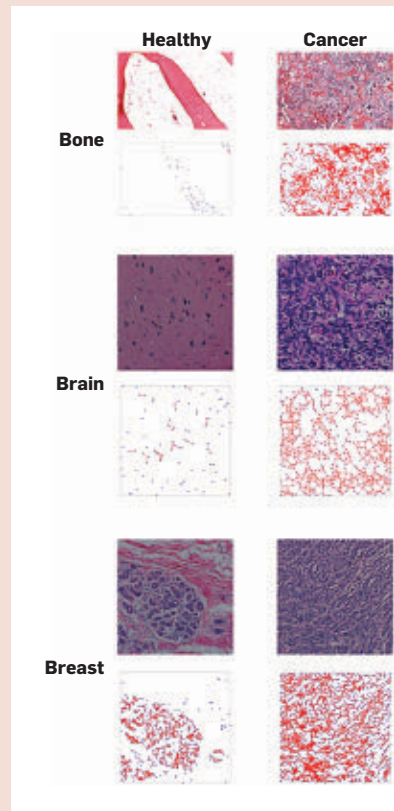
Several different approaches are used to extract features at the cellular and tissue levels to distinguish and classify distinct (mal)functional states such as tumor types in cancer. The features are used to quantify the information carried in the sample, and to distinguish the diseased structures from the healthy and damaged ones.

The first approach makes use of *morphology* to quantify the size and shape of a cell or its nucleus.⁴⁰ At the cellular level, such features are used to classify a nucleus as belonging to a healthy or diseased cell. At the tissue level, the statistics of these features over the tissue are exploited in the classification of a tissue as diseased or not.

The second approach employs *intensity*, or the distribution of the color values of pixels to define features.⁴³ Such features may include the mean, standard deviation, skewness, and kurtosis of the red, green, and blue components, as well as the difference between the red and blue components, and the proportion of the blue component in RGB color space. Given this approach directly derives features from the intensity values, these features are more sensitive to the noise that arises from stain artifacts and image acquiring conditions.

The third approach exploits the *textural descriptors* as its features and considers spatial dependency of the intensity values to quantify the smoothness, regularity or coarseness of the image. The two most popular models to compute these textural descriptors are those that use run-length matrices¹⁸

Figure 1. Different tissue types and states as well as their representations as cell-graphs.



Cell-graphs are constructed from H&E stained human tissue samples and provide precise metrics to capture the spatial organization of tissues. Using these graph metrics, machine learning algorithms can predict the functional state of underlying tissue samples. As shown in the figure, cell-graphs and corresponding functional states of different tissue types quite different.

Courtesy:doi:10.1371/journal.pone.0032227

and co-occurrence matrices.²² Fractals that describe the similarity levels of different structures found in a tissue image over a range of scales have been proposed in Einstein¹⁵ and Esgiar.¹⁶ We note that none of the approaches mentioned here can model the structure-function relationship in tissue.

Prior work using graph theory to model a tissue is based on drawing a *Voronoi graph* of cells from a tissue image.^{26,42} In these studies the graph-based features are defined on the Delaunay triangulation graph or its corresponding minimum spanning tree. There are several limitations of Voronoi graphs that cell-graphs successfully remedy. First, the edge function in Voronoi graphs is fixed and dictated

by the location of vertices. The Delaunay triangulation permits the existence of edges solely between adjacent vertices. Thus, only relationships between closely located nuclei are represented. This restriction makes it impossible to generate and test different biological hypotheses for cell-to-cell interactions. Second, Voronoi graphs are restricted to planar graphs that are very limited in their structure and do not allow crossing of edges. There is no evidence to justify such a limitation in tissue structural organization. This constraint also presents difficulties with 3D images. Third, a Voronoi graph always has a single connected component (that is, the tissue is represented by a connected graph), which may not be a valid assumption for sparse tissues (those with fewer numbers of cells). Finally, the graph features are limited and mainly computed on minimum spanning trees over Voronoi graphs.

(Various modifications have been proposed to adapt Delaunay triangulations to specific biological systems by changing the triangulation technique and resulting in different neighborhood graphs.^{2,25,37} However, the feature sets constructed from these neighborhood graphs are limited and mainly based on spanning tree properties.)

The cell-graph approach generalizes graph-based approaches by allowing an arbitrary edge function between a pair of nodes based on a biological hypothesis on their pairwise relationship. In a cell-graph, cells or cell clusters of a sample tissue are the vertices. An edge is defined between a pair of cells or cell clusters based on an assumption that has a biological foundation (or hypothesis). For example, if we believe that cells that are spatially close to each other are more likely to interact (for example, signal) with each other than more distant cells, then a link can be made between them with a probability that decays exponentially with increasing Euclidean distance between them. Thus, links of a cell-graph aim to capture the biological interactions in the underlying tissue. The cell-graphs provide a precise mathematical representation of cellular organization and the extracellular matrix (ECM) that surrounds cells. If the images carry multichannel information, by applying more sophisticated stain-

ing techniques (for example, multispectral fluorescence imaging) it is possible to build cell-graphs that have different types of nodes, corresponding to different types of cells that coexist (for example, epithelial vs. fibroblast) and other ECM entities (for example, basal membrane underlying epithelial cell layers and blood vessels). With 3D images and 3D cell-graphs, such representation becomes more accurate and powerful. Cell-graphs bring the well-established principles of graph theory and provide a rich set of features defined precisely by these principles to be used as quantitative descriptor features. These features could be defined and computed locally from a single node's point of view (for example, number of its neighbors), or globally for the entire tissue sample (that is, the shortest or longest distance in the cell-graph between any two nodes). Cell-graphs can use cell level attributes such as convexity, size, physical contact, shape, and so on to define similarity metrics for establishing links between a pair of nodes.

As an introductory example for application of the cell-graphs, consider automated diagnosis of cancer from digital images (that is, digital pathology). The “gold standard” for cancer diagnosis remains the expert (qualitative) opinion of pathologists specially trained to recognize indicative morphological signatures of different tumors in histopathology slides. This process is not only time consuming but also subject to inter-observer variability. The cell-graph method can successfully assist diagnostics by automating this process. For example, consider the problem of predicting cancer for three morphologically distinct tissues (brain, breast, and bone) from histopathology images. Figure 1 shows the cell-graphs of three different human tissue samples in two different functional states: healthy and cancerous.

Methodology

The cell-graph methodology is image-driven and utilizes different technologies ranging from fluorescence microscopy to confocal microscopy. While most of the work we report is based on hematoxylin and eosin (H&E) stained image analysis, the cell-graph technique benefits from more sophis-

ticated staining techniques discussed later in this article.

Formally, let $G=(V, E)$ denote a cell-graph with V and E being the set of nodes and edges of the graph, respectively. The overall methodology is shown in Figure 2. It starts with image analysis and ends with checking the accuracy of the machine learning algorithms.

Identification of nodes for a cell-graph. Nodes of a cell-graph are associated with individual cells, thus the first step is to distinguish the cells from their background based on the color information of the pixels. Standard imaging techniques can be used for this part of the process. We note that cell-graphs do not require precise cell segmentation and morphology since determining cell locations are enough to identify node set. Cell segmentation is an active area of research and outside the scope of this work; we refer readers to many survey papers on this topic.^{21,29,44}

One earlier approach used for node identification is to have two control parameters: the size of the grid, and the threshold value.¹¹ The grid size determines the down sampling rate, that is, the resolution of the resultant image. Consequently, a node can represent a single cell, a part of a cell, or a bunch of cells, depending on the grid size. The finer the grid size, the closer a node is to a single cell. For each grid entry, the average values of pixels located in this grid entry are computed and compared against a threshold value to determine the nodes of the cell-graph. Thresholding eliminates the noise that arises from the staining artifacts and misassignment of black pixels in the color quantization step.

There are more sophisticated approaches to cell segmentation and node identification depending on the type of the tissue and how much segmentation accuracy is required. For example, cells can be identified by us-

Figure 2. Methodology for image-driven tissue modeling.

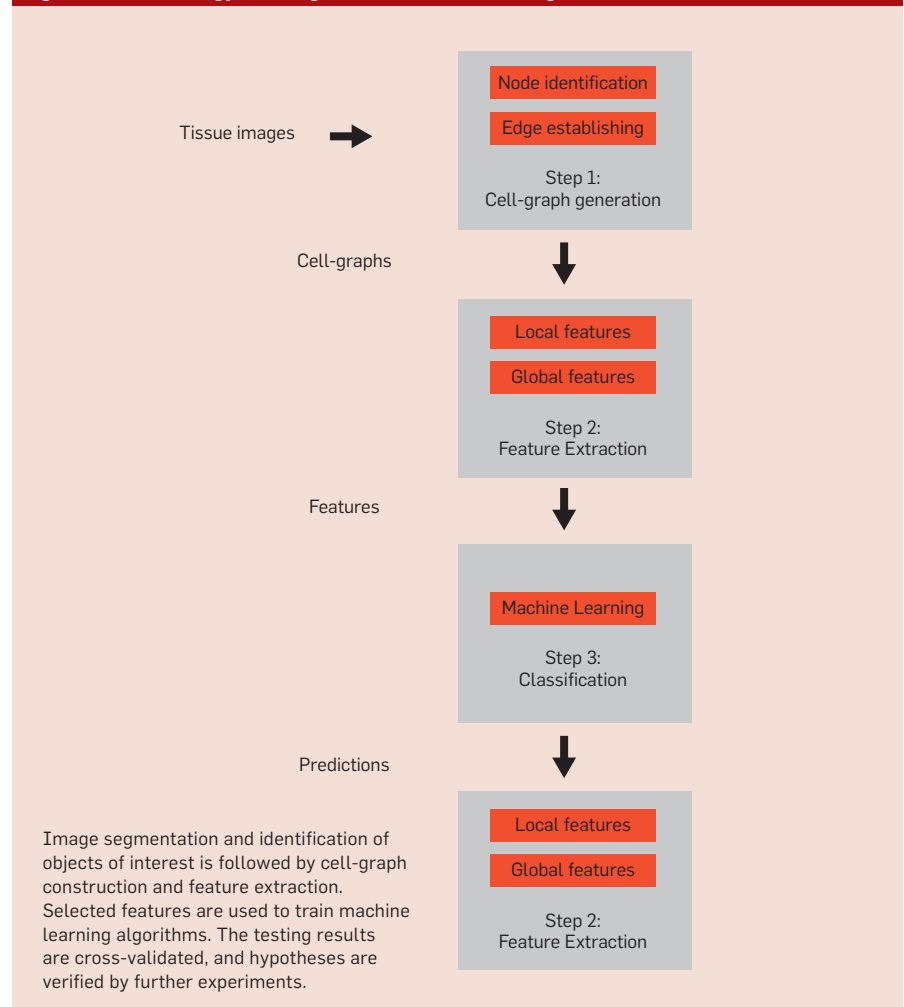
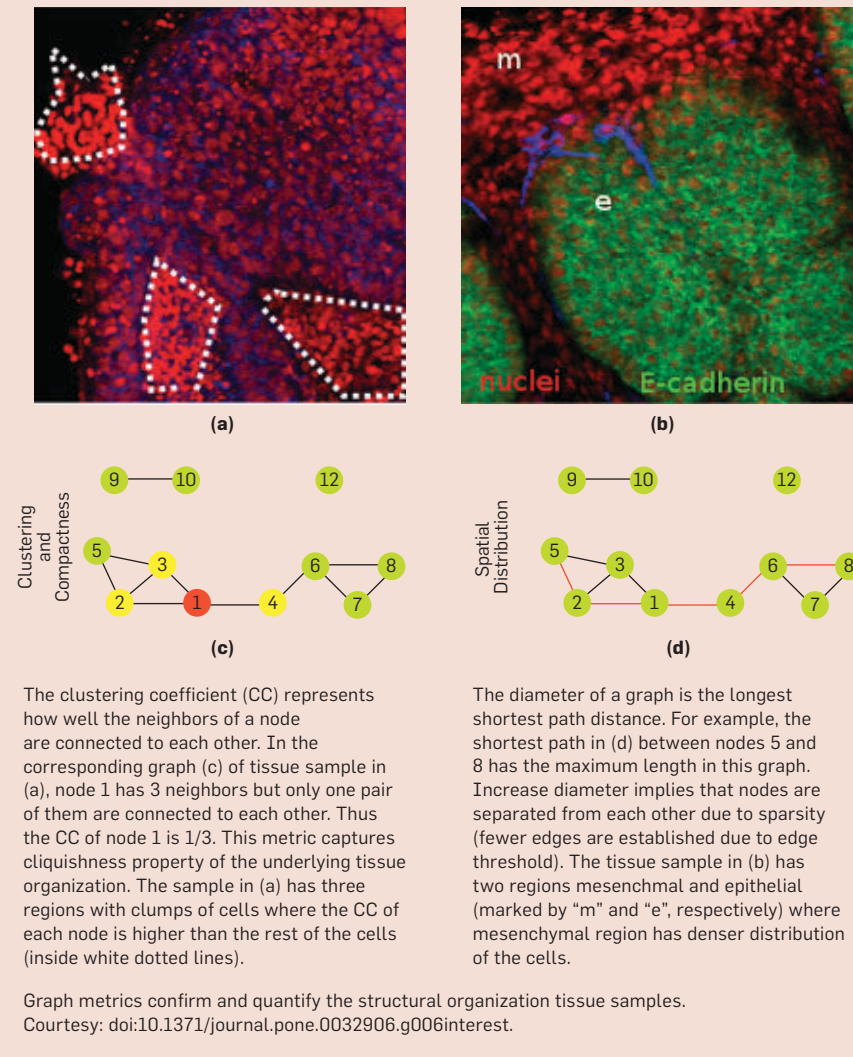


Figure 3. Geometric interpretation of changes in cell-graph features.

ing eigenvalues of the Hessian matrix of the image^{17,23} with two parameters of interest: $R_b = \lambda_1/\lambda_2$, and $S = \|\nabla^2 f\|$ where S is the Frobenius matrix norm and is used to differentiate objects of interest from the background, whereas R_b is a measure to differentiate between blob-like structures and ridge-like structures. S will be low in background pixels as the eigenvalues for pixels lacking contrast will be small. In high-contrast regions however, at least one of the eigenvalues will be high and S will be large.⁴ For further details, we refer the reader to several excellent survey papers on this topic as cited above.

Establishing edges in a cell-graph. After determining the vertex set V , an edge (u, v) between a pair of nodes u and v can be defined by making use of the biological insight and knowledge of the interaction of the cells in a spe-

cific tissue type. For example, it may be more likely that physically adjacent cells signal each other than the ones far away. Such distance-based interaction among the elements is well understood in physical systems based on energy minimization. In the absence of any other similarity measure between a pair of cells, one can adapt a simple Euclidean distance measure for defining an edge between them. Therefore, we translate the pairwise spatial relation between every two nodes to the possible existence of links in a cell-graph.

An edge (u, v) can be established probabilistically or deterministically or a combination of these two methods. For example, in probabilistic cell-graphs¹¹ the probability of creating a link between any two nodes may decay exponentially with the Euclidean distance between them employing a

probability function $P(u, v) = e^{d(u, v)/(L)}$ where $d(u, v)$ is the Euclidean distance, and L is the largest Euclidean distance between two nodes of the grid. The model parameters α and β must be chosen between 0 and 1. These parameters affect the number of the links and the connectivity of the graphs. Selecting smaller values of these parameters results in a smaller number of links. Different probability functions such as power law $P(u, v) = d(u, v)^{-\alpha}$ can also be used based on a different hypothesis. Intuitively, the closer two cells are, the more likely that they share a relationship. This probability quantifies the possibility for one of these nodes to be grown from the other, thus aiming to model the prevalence of the disease state in a tissue.

An edge (u, v) can be established deterministically if the distance $d(u, v)$ is less than a threshold (for example, two cells are physically touching each other). This edge function captures cell interaction through adhesion. When the dataset is large, cross-validation techniques can be used to identify the optimal threshold that might signify cell-cell communication. In cases when the dataset is limited in size, heuristics such as five times the average radius of a nucleus (for example, 20 microns) can be used.

Note that the presence of a link between nodes does not specify what kind of relationship exists between the nodes (cells); it simply indicates that a relationship of some sort is hypothesized to exist, and that it is dependent on the distance between cells. Surprisingly, the distance measure alone is sufficient to reveal important, diagnostic structural differences in human tissues (see sidebars 1–3 in the online appendix).

Feature extraction. After constructing the cell-graphs, the next step is to define and extract graph features to train machine learning algorithms for classification of tissue functional states. We consider two types of features to be used by classification algorithms: local features at the individual cell level, and global features at the tissue level. Table 1 (in an online appendix accompanying this article in the ACM Digital Library dl.acm.org) summarizes the graph features to capture information from different scales.⁶ By computing the distri-

bution of local features, one can obtain some of the global features. However, some other global features, such as the ratio of the size of the giant connected component over the size of the entire graph, can only be computed over the entire graph.

The spectrum of a graph, which is the set of graph eigenvalues computed from the adjacency matrix or from its Laplacian, also provides global features such as the spectral radius and Eigen exponents. The eigenvalues of the Laplacian relate to the graph invariants better than the eigenvalues of the adjacency matrix.⁸ For example, the number of eigenvalues with a value of 0 gives the number of connected components in the graph. Moreover, as the eigenvalues of the Laplacian lies in the range $[0,2]$, it is easier to compare the spectra of graphs with different sizes.

Feature selection and machine learning. Feature selection helps to overcome the problem of curse of dimensionality and may increase classification accuracy. Note the importance of graph features vary from one type of tissue to another. For example, the most important features for bone tissue classification (using f-scores for feature selection) are the *number of nodes* (fs=1.685), *giant connected ratio* (fs=1.094), *number of central points* (fs=1.607), and *clustering coefficient* (fs=1.069) (the next f-score value corresponds to *percentage of end points* which is much smaller).⁴ Interestingly, while the number of nodes (that is; cell density) is an important feature for brain tissue analysis, it is not so for breast tissue.⁵ Some features such as *clustering coefficient* and *number of central points* are important for bone tissue, as well as for breast and brain tissue. Other influential features include *average effective eccentricity*, *number of links*, and *average path length*. While there is a tremendous amount of work on it (for an online repository see <http://featureselection.asu.edu/>), feature selection is based on heuristics and different methods yield different subsets.

The selected graph features are input to a classifier such as the Artificial Neural Networks, Bayesian Networks, or Support Vector Machines (SVM) for learning and predicting the functional state associated with the structure. The main challenge from learning perspec-



The cell-graph technique aims to learn structure-function relationship by modeling structural organization of a tissue/organ sample using graph theory.



tive is the unbalanced class representation. For example, while there is abundant data labeled as “cancer” class for almost all the tissue types we studied, much less data was available labeled as “healthy” class. Standard techniques such as under sampling and over sampling of data are applied to cope with this problem, and in addition each dataset is normalized and centered. In SVM the test data is classified by determining the side of the hyperplane they lie on in the kernel-mapped space. The radial basis function (RBF) kernel, also referred to as the Gaussian kernel (that is, $K(x_i, x_j) = \exp(-||x_i - x_j||^2 / 2\sigma^2)$), is commonly used as a kernel to map the data into an infinite dimensional Hilbert space. While there are some parameters in SVM that can be fine-tuned to increase learning accuracy, the default settings used in Matlab, LibSVM, and other packages have shown to be sufficient. In order to extend SVM for the classification of three classes, one can employ the one-against-one approach²⁴ where three two-class SVM classifiers are established for each pair of classes in the training dataset. Each sample in the test data is assigned to a class by these classifiers and the class with the majority vote is chosen as the final result. If there is equal voting for the three classes, the class that has the largest margin from the separating hyperplane is chosen. The Bayesian classifier maximizes the posterior probability, which is a function of the likelihood and prior probability with the assumption that the data points are drawn from Gaussian distributions. The KNN (K-nearest neighborhood) classifies each data point to the class that is most common among its K-nearest neighbors determined by a Euclidean distance-based difference. In this study, we test three values $K=10, 11$, and 12 , and choose the values that achieve the highest grading accuracy. Both Bayesian and KNN classifiers can readily handle multiclass classification problem.

We note that classification accuracy of different machine learning algorithms may vary on the same feature set. For example, for histopathological grading of follicular lymphoma (FL) images into one of three grades, a comparison of three classifiers (SVM, Bayesian, and KNN) show different accuracy results³⁴ (see Table 3 in the online appendix).

Finally, the cell-graph features can be used to design specialized kernels as an alternative approach to RBF or polynomial kernels for graph classification problems. Such kernel computation is based on feature-vectors constructed from different global topological attributes, as well as global label features. The main idea²⁴ is the graphs from the same class should have similar topological and label attributes. A detailed comparison on real benchmark datasets shows that our topological and label feature-based approach delivers better or competitive classification accuracy, and is also substantially faster than other graph kernels.²⁷ It is the most effective method for large unlabeled graphs.

Feature interpretation. There are three types of cell-graph features: cliqueness metrics (for example, clustering coefficient), compactness metrics (for example, number of central points), and distance metrics (for example, diameter).

It is best to explain the meanings of graph features within an application domain. Recently, the cell-graph technique was used to quantify changes in the cellular dynamics of submandibular gland (SMG) morphogenesis as a function of ROCK1-mediated signaling in this process.⁶ The laboratory analysis verified that the average diameter of the SMG increased and that the thickness decreased following inhibitor treatment, which is consistent with the overall decrease in cellular contractility (see Figure 3). Additionally, the total number of cells is also decreased with inhibitor treatment. This implies the overall compactness of the explant decreases both at the tissue and at the cellular level with ROCK inhibitor-treatment. The values for certain cell-graph features captured this observation: the clustering coefficient in the control tissues was greater than in the ROCK inhibitor-treated tissues. The clustering coefficient gives a measure of compactness of a tissue. That is, cells in the

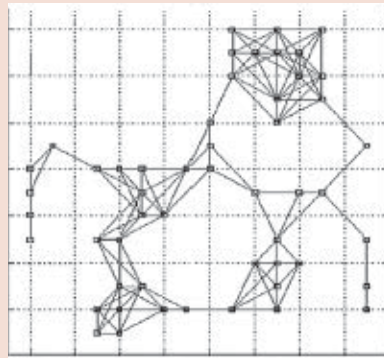
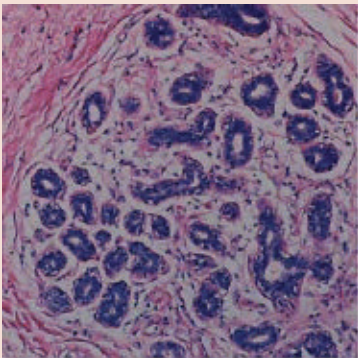
ROCK inhibitor-treated tissues were further apart from each other and thus, had fewer edges, or links, per unit area, measurable as a decreased clustering coefficient. The average path length, which measures the average shortest path between two cells, increases with ROCK inhibition and number of connected components, which is the number of cell-linked cell clusters, decreases. Less compact tissue should have a smaller number of linked cells, an increased inter-cellular distance (that is, longer average path length) and, hence, a lower number of connected components. Cell-graph features were thus able to predict known ROCK inhibitor-induced global tissue changes.⁶

Enhanced Cell-Graph Models

This section explores how to go beyond simple cell-graphs without self-loops, multiple edges, and attributes to more complex ones.

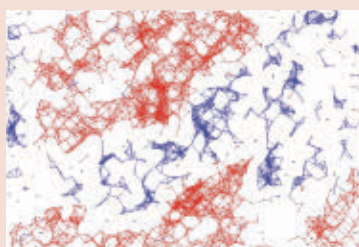
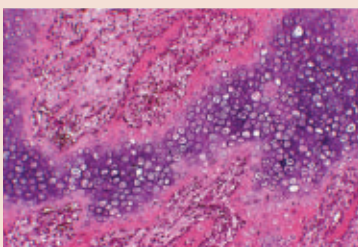
Hierarchical cell-graphs. So far the cell-graphs discussed are used to model diffusive structural organizations such as the ones found in brain tissue samples. Other tissue types such as breast or prostate exhibit more complex structural organizations and require enhancing the cell-graph approach further⁵ (see comparison results in Sidebar 2 in the online appendix). For example, to model the lobular structure of breast tissue, a 2-phase cell-graph construction is proposed⁵ (see Figure 4). After cell segmentation, first, connected subgraphs have been built to capture the local structure of lobular/glandular architecture so that each connected component represents a lobular/glandular structure. A biologically meaningful hypothesis for this step is that within a glandular structure there is a high interaction among the cells as a function of physical contact. Second, the interactions among the connected components are modeled with the hypothesis that the likelihood of interglandular interaction through ECM may decrease as a function of the spatial distance between them. These two phases may admit different edge functions. For example, intra-glandular edges can be assigned deterministically while interglandular edges would be defined probabilistically.⁵ The hierarchical cell-graphs enable us to model and test different biological hypoth-

Figure 4. Hierarchical cell-graphs for breast tissue modeling.



Each connected component of a hierarchical cell-graph corresponds to a lobular structure, which is modeled by using simple cell-graph approach.²³ Courtesy: Conf Proc IEEE Eng Med Biol Soc. 2007;2007:5311-4.

Figure 5. ECM aware cell-graphs.



Fractured bone tissue example where fracture cells exist in the middle of the original image and stem cells repairing fracture are colored in blue. Courtesy: doi: 10.1007/s10618-009-0153-2

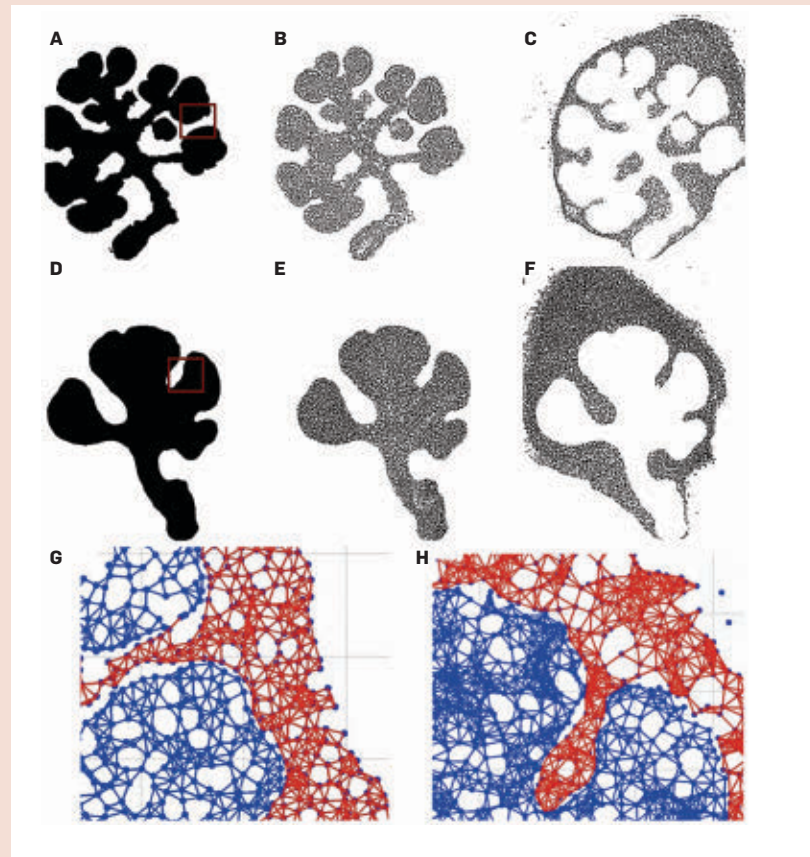
esizes on the interaction of cells, and glands by changing the edge function.

ECM-aware cell-graphs. In a tissue sample, there is more than one type of cells of interest, including blood cells, cancerous cells, normal cells, stem cells repairing a damaged tissue (for example, bone fracture), and so on. These cells are not only distinguishable from each other by their color and size, but also carry valuable information about the underlying functional state. ECM-aware cell-graphs take advantage of the heterogeneity of tissue samples to encode more information. After image segmentation, each cell is assigned a color-code based on the ECM composition of its surroundings. For each color code, a dedicated cell-graph is constructed and graph features are extracted. As a result, multiple cell-graphs coexist for modeling the same tissue and their combined feature set can be used for tissue classification. Figure 5 shows a tissue sample from fractured bone and corresponding cell-graphs constructed from segmentation, which illustrates that ECM-aware cell-graphs result in high accuracy in bone tissue classification problems⁴ (see Sidebar 3 online).

Similarly, Figure 6 captures the spatial organization of epithelial, and mesenchymal cells that coexist during the branching morphogenesis of submandibular gland.⁶

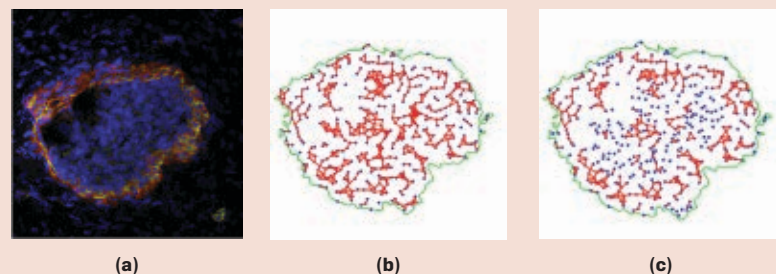
Cell-graphs with multiple staining. With immunohistochemical staining becoming more widely used in digital pathology practices, richer sets of biological information also become available to construct cell-graphs that are more realistic and relevant. These staining techniques indicate the expression level of various proteins and provide important information for learning the underlying functional state. For example, for the automated grading of breast cancer in 3D tissue sections, both the distribution and expression level of a lateral cell-membrane protein integrin $\alpha 3$ has been used to hypothesize the interaction between cell pairs.³⁵ The cell-graph edges have been established based on the integrin $\alpha 3$ densities between the nuclei pairs (see Figure 7). As the cancer progresses, reduced expression of integrin $\alpha 3$ is observed corresponding to the loss of interaction between the cells which is an important feature for

Figure 6. Stitched images of submandibular gland were segmented using the active contour method to define epithelial (white) vs. mesenchymal tissue (black) in control (a) and ROCK inhibitor-treated explants (d).



These masks were used to identify the epithelial nuclei (b, e) and mesenchymal nuclei (c, f). Using each nucleus as a vertex, cell-graphs were constructed for control and ROCK inhibitor-treated tissues, respectively (g, h), where zoomed regions of cell graphs corresponding to regions of the original images (shown as red boxes in a and d) are shown in detail. Epithelial tissue is represented by the blue graph and the mesenchymal tissue is represented by the red graph. We discarded the sublingual tissues and only used the submandibular gland. Courtesy: doi: 10.1371/journal.pone.0032906.g002

Figure 7. Non-invasive breast tissue sample.



Non-invasive breast tissue sample with two stains is shown in (a). Corresponding cell-graph based on 3D distance is shown in (b) and the cell-graph that considers expression levels is shown in (c). Table 4 in appendix shows the improvement in classification by capturing the underlying biology more effectively. Courtesy: doi: 10.1109/ISBI.2013.6556431.

grading. Table 4 in the online appendix illustrates that grading accuracy increases by capturing this information.

Recently, a new technique called *hyperplexed immunofluorescence technology* has allowed unlimited stains on a single specimen.¹⁹ Molecular stains are quantified in single cells and sub-cellular compartments, yielding unparalleled insights into the biology of intact tissues. Each stain reflects a different biological property—cell types, signaling processes, and so on. As an extension of ECM-aware cell-graphs, one can establish links based on spatial proximities of pairs of cells using the expression of each marker, thus building one cell graph for each marker. We expect the feature set of each cell-graph captures a different biological property. Figure 8 shows the cell-graphs for three different stains. Statistical analysis shows that feature sets obtained from these cell-graphs come from different probability distributions (see Table 5 in the online appendix).

3D tissue analysis with cell-graphs.

3D confocal imaging techniques have been a powerful tool for cell biologists and engineers providing 3D spatial information regarding the location of specific structures within cells and tissues. Some of the work discussed previously used 3D modeling.^{6,35} To capture the 3rd dimension, one needs to stitch z-stack sequences with some overlapping. This process requires defining a depth parameter k and then ensuring some overlap between stack $i-1$, stack i , and stack $i+1$. For example, in Oztan et al.³⁵ z-stack sequences of 8 slices deep with 25% overlap with the preceding and following z-stack sequences

are constructed. It is straightforward to extend 2D Euclidean distance to 3D. Consider two vertices: $u = (x_u, y_u, z_u)$ and $v = (x_v, y_v, z_v)$ then the distance between them in 3D would be

$$\sqrt{(x_u - x_v)^2 + (y_u - y_v)^2 + (z_u - z_v)^2}$$

Similarly, different distance metrics can be adapted such as the Lp distance:

$$L_p = \left(\sum_{i=1}^3 |x_i - y_i|^p \right)^{1/p}$$

The features defined on 2D cell-graphs can be used in the 3D case with additional computational demand.

3D cell-graphs have been constructed to model a 3D cellular environment and quantify type I collagen remodeling and fibrillogenesis with respect to mesenchymal stem cell organization over time.³ In that work, an initial result on how to integrate a physics-based mechanical model³⁰ with the cell-graph approach is also shown. The results are verified on multiple experiments and provide the first quantitative support to the hypothesis that continuity between extracellular and intracellular environments is required for stem cell fate determination (see Figure 9 in the online appendix).

Time Series of Tissue Evolution

Up to this point, the discussion on cell-graphs was confined to static histology samples. Here, we are interested in modeling the evolution of time dependent cell and tissue growth.

Spatiotemporal cell differentiation.

Recently, in vitro (3D hydrogel models) evolutions of 11 different cell-lines from different tissues that develop solid tu-

mors (epithelial, connective, and neural) were studied in order to determine which structural properties (captured by graph metrics) dominate the differentiation.²⁸ The cell-lines include MCF10A: Precancerous Human, Breast Epithelial; AU565: Human Breast Cancer HER2+/ER-; MCF7: Human Breast Cancer HER2-/ER+; MDA-MB231: Human Breast Cancer HER2+/ER2+; hDFB: Human Dermal Fibroblasts; NHA: Normal Human Astrocytes; U118MG: Human Glioblastoma; NHOst: Normal Human Osteoblast; MG63: Human Osteosarcoma; RWPE-1: Non-tumorigenic Human Prostate; DU145: Human Prostate Carcinoma.²⁸ Although there are limitations to in vitro studies, the cell lines used in this study represent a range of tissue types allowing one to directly compare the structural profiles of various functional states through analysis of cell-graph metrics as follows.

The structural features of underlying tissue samples are calculated on each cell-graph using $G_i = (V_i(t), E_i(t))$, where $V_i(t)$ and $E_i(t)$ represent the list of vertices and nodes at time point t and i represents the index for the cell line. As a result, time series of tissue evolution can be represented by a 3rd order tensor with the modes: *features* $\times$ *time* $\times$ *cell-line* whose dimensions are I, J , and K , respectively.³⁰ An entry T_{ijk} in this data cube corresponds to the value of metric i at time point j for cell-line k where $i = 1, \dots, 20; j = 1, \dots, 6$; and $k = 1, \dots, 11$.

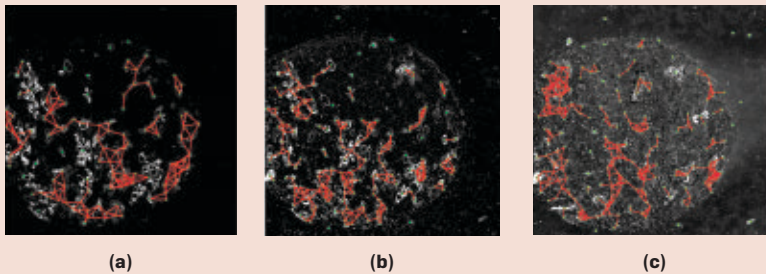
Two common models in multi-way data analysis are Tucker3 and Parallel Factor Analysis (PARAFAC).¹ A Tucker3 model with orthogonality constraints on component matrices is a generalization of SVD from matrices to high-order datasets and is also called Higher-Order Singular Value Decomposition (HOSVD)¹⁰ or multilinear SVD. Using a Tucker3, a 3-way tensor $T \in \mathbb{R}^{I \times J \times K}$ is modeled as follows:

$$T_{ijk} = \left(\sum_{r=1}^R \sum_{q=1}^Q \sum_{p=1}^P G_{pqr} A_{ip} B_{jq} C_{kr} \right) = E_{ijk}$$

where P, Q and R indicate the number of components extracted from first, second and third mode ($P \leq I, Q \leq J$, and $R \leq K$), respectively. $A \in \mathbb{R}^{I \times P}$, $B \in \mathbb{R}^{J \times Q}$, and $C \in \mathbb{R}^{K \times R}$ are the component matrices. $G \in \mathbb{R}^{P \times Q \times R}$ is the core tensor and $E \in \mathbb{R}^{I \times J \times K}$ represents the error term.⁵

Three-mode tensor analysis in feature mode for outliers identified six features such as; average degree; clus-

Figure 8. Cell-graphs overlaid on images of (a) E-Cadherin (b) Pan-Keratin (c) Keratin 15.



The cell-graphs reveal distinct spatial patterns showing orthogonal information can be obtained.

tering coefficient; number of central points; number of connected components; standard deviation of edge lengths; and number of isolated points that capture the *compactness*, *clustering*, and *spatial uniformity* of the 3D architectural changes for each cell type throughout the time course.²⁸ Importantly, four of these metrics are also the discriminative features for our histopathology data from the previous studies reviewed earlier.

Dynamic cell-graphs. Spatiotemporal development of tissues/organs requires modeling of cell-to-cell interactions over time and has proven to be difficult. For example, while the branching processes in developing organs (lungs, pancreas, kidneys, salivary, and mammary glands) have been studied in detail, we are still far from comprehending the integrated process.⁹ Computational modeling of morphogenesis starts with mathematical models for understanding the fundamental properties of cell clusters.^{14,41} These theories were followed by continuum, physics-based models, which considered a tissue to be composed of cells and ECM and described the stress forces between these two structures.^{32,33} Such models found a wide area of applications including modeling of epithelial morphogenesis in 3D breast culture acini,³⁸ as well as lung³¹ and kidney branching morphogenesis.³⁹ However, these models are data agnostic and focus on optimization of the model parameters for the best outcome.

Advanced imaging techniques provide a vast amount of image data that motivates data-driven modeling approaches. Data-driven techniques based on cell-tracking (identifying the same cell over different time points) are computationally challenging at the organ level. The cell-graphs provide a scalable alternative by tracking the graph properties instead of individual cells. For example, dynamic cell-graphs have been constructed to model the growth and cleft formation in SMG branching morphogenesis.⁴⁵ The model takes the initial gland morphology and nuclei locations from an initial image along with basic biological control parameters such as epithelial growth factor (EGF) concentration as input. The EGF concentration levels determine the mitosis and cleft deepening rates. At each



Data-driven techniques based on cell-tracking are computationally challenging at the organ level. Cell-graphs provide a scalable alternative by tracking the graph properties instead of individual cells.



iteration of the algorithm, the cells are divided into two populations based on their distance from the gland boundary, namely internal (I) and periphery (P). Subsets I_0 and P_0 of I and P , respectively are chosen to undergo a proliferation attempt. Cells in P_0 that successfully undergo mitosis create new cells (or vertices) V_0 that are added to V .

Cells with identical topology and growth are permitted only at the gland boundary, where a hypothesized “nutrient medium” provided by the mesenchyme is accessible. In the dynamic cell-graph model, this similarity is enforced via the local structural (graph) properties of cell-graphs that maintain consistency in the topology of the SMG throughout the development stages. When first created, potential daughter vertices are placed outside the initial gland boundary in a region within 20° of the surface normal from the parent vertex at a minimum distance of one cell diameter, but less than the specified maximum edge length. Some parameter K of possible candidate daughter vertices satisfying these spatial and angular constraints are chosen, and the daughter vertex with the closest local cell-graph features to the parent vertex is selected as the optimal daughter vertex. These local structural features assess the spatial uniformity (clustering coefficient), connectedness (degree, closeness centrality, betweenness centrality), and compactness (edge length statistics) of the cell-graph. New edges, E_0 , are also constructed based on the distances from the new cells to existing cells in G . Bud outgrowth is modeled by the annexation of new nodes into the gland boundary.³³ However, the main difficulty with this approach is to derive the smooth shape formation from the cell-to-cell interactions as discussed in Dhulekar et al.¹³

Conclusion

This article explored various cell-graph constructions to model the information encoded in the image of a complex structure like the human brain or bone tissue. The main assumption of the cell-graph approach is that cells in a tissue organize in a certain way to perform a specific function; thus, understanding structure would predict the function or malfunction.

Graph theory implementation pro-

vides a rich and rigorous set of features that are almost an order of magnitude greater than prior methods, which were limited to a few features. Furthermore, it facilitates new kernels and data mining algorithms such as subgraph mining. These features can then be used to train machine learning algorithms for predicting the functional class label, given the feature set for test data. We note that identifying graph metrics that help to predict long-term functionality by linking engineered tissue structure to function is an important step toward optimizing biomaterials for the purposes of regenerative medicine.

Any interdisciplinary work requires strong collaboration between biomedical experts and computational scientists, given that interpretability of the results is crucial. It is particularly important to understand and relate the computational feature space back to the original problem domain to advance the knowledge there. Some of the cell-graph features are not intuitive while still useful for classification and prediction (while they remain effective, interpretability of features poses a specific challenge to the convoluted features obtained by deep learning algorithms).

Finally, we note that a coupling of continuum physics-based models with discrete data-driven models such as the cell-graphs may provide more accurate prediction as the complexity of the underlying problem increases (for example, organ morphogenesis). In Metzger et al.³¹ an initial attempt for such model combining is reported by replacing the “springs”³² with weighted cell-graph edges where weights are calculated directly from images of collagen fibers. However, much work needs to be done in this direction since as reported,⁴⁵ the cell-graphs are agnostic to the physical laws that govern the underlying structural organization, and are sufficient to predict complex shape formation and need to be coupled with techniques such as the level set method. 

Additional background information, literature, and figures appear in an online appendix available with this article in the ACM Digital Library (<http://dl.acm.org/citation.cfm?id=2960404&picked=formats>).

References

- Acar, E. and Yener, B. Unsupervised multiway data analysis: A literature survey. *IEEE Transactions on Knowledge and Data Engineering* 21, 1 (2009), 6–20.
- Albert, R., Schindewolf, T., Baumann, I. and Harms, H. Three-dimensional image processing for morphometric analysis of epithelium sections. *Cytometry* (1992), 13:759–765.
- Bilgin, C.C. et al. Quantification of three-dimensional cell-mediated collagen remodeling using graph theory. *PLoS One* 5, 9 (2010), e12783.
- Bilgin, C.C., Bullough, P., Plopper, G.E., and Yener, B. ECM-aware cell-graph mining for bone tissue modeling and classification. *Data Min. Knowl. Discov.* 20, 3 (May 2010), 416–438.
- Bilgin, C., Demir, C., Nagi, C. and Yener, B. Cell-graph mining for breast tissue modeling and analysis. In *Proc. of IEEE EMBC* (2007).
- Bilgin, C.C., Ray, S., Baydil, B., Daley, W.P., Larsen, M. and Yener, B. Multiscale feature analysis of salivary gland branching morphogenesis. *PLoS One* 7, 3 (2012).
- Car, R. and Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Physical Review Letters* 55, 22 (1985), 2471–2474.
- Chung, F.R.K. Spectral graph theory. Conference Board of the Mathematical Sciences, *American Mathematical Society* 92 (1997). Providence, RI.
- Davies J. *Branching Morphogenesis*. Springer-Verlag, 2004.
- de Lathauwer, de Moor, L.B. and Vandewalle, J. A multilinear singular value decomposition. *SIAM J. Matrix Analysis and Apps* 21, 4 (2000), 1253–1278.
- Demir C., Gultekin, S.H. and Yener, B. Augmented cell-graphs for automated cancer diagnosis. *Bioinformatics (Suppl 2)* 21, (2005), ii7–ii12.
- Demir C., Gultekin, S.H. and Yener, B. Learning the topological properties of brain tumors. *IEEE/ACM Trans. Computational Biology and Bioinformatics* 2, 3 (2005), 262–270.
- Dhulekar, N., Oztan, B. and Yener B. Model coupling for predicting a developmental patterning process. In *Proc. of SPIE* 2016.
- Eden, M. A two-dimensional growth process. In *4th Berkeley Symposium on Mathematical Statistics and Probability* (1961), 223–239.
- Einstein, A.J., Wu, H.S., Sanchez, M. and Gil, J. Fractal characterization of chromatin appearance for diagnosis in breast cytology. *Journal of Pathology* 185, 4 (1998), 366–381.
- Esgiar, A.N., Naguib, R.N.G., Sharif, B.S., Bennett, M.K., Murray, A. Fractal analysis in the detection of colonic cancer images. *IEEE Trans. Information Technology in Biomedicine* 6, 1 (2002), 54–58.
- Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A. Multiscale vessel enhancement filtering. *Lecture Notes in Computer Science* (1998), 130–137.
- Galloway M.M. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*. (1975) 4:172–179.
- Gerdes et. Al. Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *PNAS* 110, 29 (2013), 11982–11987.
- Gunduz, C., Yener, B., and Gultekin, S.H. The cell graphs of cancer. *Bioinformatics* 20 (2004), 1145–1151.
- Gurcan, M.N., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N. and Yener, B. *Histopathological Image Analysis: A Review*. 2009.
- Haralick R.M. Statistical and structural approaches to texture. In *Proc. of IEEE*. 67, 5 (1979), 786–804.
- Hladuvka, J., Konig, A. and Grotler, E. Exploiting eigenvalues of the Hessian matrix for volume decimation. In *Proceeding of the 9th International Conference in Central Europe on Computer Graphics, Visualization, and Computer Vision* (2001), 124–129.
- Hsu, C. and Lin, C. A comparison of methods for multiclass support vector machines. *IEEE Trans. on Neural Networks* 13, 2 (2002), 415–425.
- Jaromczyk J.W. and Toussaint G.T. (1992). Relative neighborhood graphs and their relatives. In *Proc. IEEE* 80 (1992), 1502–1517.
- Keenan S.J., Diamond, J., McCluggage, W.G., Bharucha, H. Thompson, D., Bartels, B.H. and Hamilton, P.W. An automated machine vision system for the histological grading of cervical intraepithelial neoplasia. *J. Pathol.* 192, 3 (2000), 351–362.
- Li, G., Semerci, M., Yener, B. and Zaki, M.J. Effective graph classification based on topological and label attributes. *ASA Data Science J. Statistical Analysis and Data Mining* 5, 4 (2012), 265–283.
- McKeen-Polizzotti, L. et al. Quantitative metric profiles capture three-dimensional temporospatial architecture to discriminate cellular functional states. *BMC Medical Imaging* 11.1 (2011), 1.
- Meijering, E. Cell segmentation: 50 years down the road. *IEEE Signal Processing Magazine* 29, 5 (Sept. 2012), 140–145.
- Meineke, F., Potten, C. and Loeffler, M. Cell migration and organization in the intestinal crypt using a lattice-free model. *Cell Proliferation* 34 (2001), 253–266.
- Metzger, R., Klein, O., Martin, G. and Krasnow M. The branching programme of mouse lung development. *Nature* 453 (2008), 745–750.
- Murray, J.D. and Oster, G.F. Generation of biological pattern and form. *Math Med Biol.* 1 (1984), 51–75.
- Oster, G.F., Murray, J.D., and Harris, A.K. Mechanical aspects of mesenchymal morphogenesis. *J. Embryol Exp Morphol* 78 (1983), 83–125.
- Oztan, B., Kong, H., Gurcan, M.N. and Yener, B. Follicular lymphoma grading using cell-graphs and multi-scale feature analysis. *SPIE Medical Imaging*. International Society for Optics and Photonics, 831516–831516.
- Oztan, B., Shubert, K.R., Bjornsson, C.S., Plopper, G.E. and Yener, B. Biologically-driven cell-graphs for breast tissue grading. In *Proceedings of IEEE 10th International Symposium on Biomedical Imaging (Apr. 2013)*, 137–140.
- Plopper, G., Larsen, M. and Yener, B. *Image-enhanced Systems Biology: A Multiscale, Multidimensional Approach to Modeling and Controlling Stem Cell Function in Computational Biology of Embryonic Stem Cells*. Ming Zhan, ed. Bentham Science Publishers, 2012, 71–87.
- Raymond, E., Raphael, M., Grimaud, M., Vincent, L., Binet, J.L., Meyer, F. Germinal center analysis with the tools of mathematical morphology on graphs. *Cytometry* 14 (1993), 848–861.
- Rejniak, K.A. An immersed boundary framework for modeling the growth of individual cells: An application to early tumour development. *J. Theor Biol* 247, 1 (2007), 186–204.
- Srivathsan, A., Menshikau, D., Michos, O. and Iber, D. Dynamic image-based modelling of kidney branching morphogenesis. *Computational Methods in Systems Biology. Lecture Notes in Computer Science* 8130 (2013), 106–119. Springer, Berlin Heidelberg.
- Street W.N., Wolberg, W.H. and Mangasarian, O.L. Nuclear feature extraction for breast tumor diagnosis. *IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*. San Jose, CA, 1905:861–870.
- Turing A.M. The chemical basis of morphogenesis. *Philos Trans R Soc Lond B Biol Sci* 237, 641 (1952), 37–72.
- Weyn, B. et al. Computer-assisted differential diagnosis of malignant mesothelioma based on syntactic structure analysis. *Cytometry* (1999), 35:23–29.
- Wiltgen M., Gerger, A. and Smolle, J. Tissue counter analysis of healthy common nevi and malignant melanoma. *Int J Med Inform.* 69(1), 17–28, 2003.
- Xing, F. and Yang, L. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review. *IEEE Rev Biomed Eng.* (Jan. 6, 2016).
- Yener, B., Dhulekar, N., Ray, S., Yuan, D., Oztan, B., Baskaran, A. and Larsen, M. Prediction of growth factor dependent cleft formation during branching morphogenesis using a dynamic graph-based growth model. *IEEE/ACM Trans. Computational Biology and Bioinformatics*.

Bülent Yener (yener@cs.rpi.edu) is a professor in the Department of Computer Science and in the Department of Electrical, Computer and Systems Engineering at Rensselaer Polytechnic Institute, Troy, NY. He is the founding director of Data Science Research Center at RPI as well as co-director of Pervasive Computing and Networking Center.

Copyright held by owner/author.



Watch the author discuss his work in this exclusive *Communications* video. <http://cacm.acm.org/videos/cell-graphs>

research highlights

P. 86

Technical Perspective Magnifying Motions the Right Way

By Richard Szeliski

P. 87

Eulerian Video Magnification and Analysis

By Neal Wadhwa, Hao-Yu Wu, Abe Davis, Michael Rubinstein, Eugene Shih, Gautham J. Mysore, Justin G. Chen, Oral Buyukozturk, John V. Guttag, William T. Freeman, and Frédo Durand

P. 96

Technical Perspective Mapping the Universe

By Valentina Salapura

P. 97

HACC: Extreme Scaling and Performance Across Diverse Architectures

By Salman Habib, Vitali Morozov, Nicholas Frontiere, Hal Finkel, Adrian Pope, Katrin Heitmann, Kalyan Kumaran, Venkatram Vishwanath, Tom Peterka, Joe Insley, David Daniel, Patricia Fasel, and Zarija Lukić

Technical Perspective

Magnifying Motions the Right Way

By Richard Szeliski

THE ABILITY TO reliably amplify subtle motions in a video is a wonderful tool for investigating a wide range of phenomena we see in the natural world. Such techniques enable us to visualize the subtle blood flow in a person's face, the rise and fall of a sleeping infant's chest, the vibrations of a bridge swaying in the wind, and even the almost imperceptible trembling of leaves due to musical notes.

The development of image processing techniques to amplify such small motions is one of the recent breakthroughs in the computational photography field, which applies algorithmic enhancement techniques to photos and videos in order to create images that could not be captured with regular photography. Some of the earlier work on this topic (originating from the same research group at MIT) used motion estimation (optical flow) techniques to recover small motions, amplify them, and then digitally warp the images. Unfortunately, optical flow techniques are very sensitive to noise, lack of texture, and discontinuities, which make this approach very brittle.

More recently, the idea of adding scaled amounts of temporal intensity differences, which the authors call the *Eulerian approach* because of its connection to fluid dynamics (which also models motion), has produced a simpler—and in many cases more robust—approach. However, this technique also amplifies noise, and it breaks down for larger amplification factors.

To see why this is the case, think of a thin line (say a telephone wire) swaying slightly in the wind. The main difference between two adjacent video frames is a darkening of the sky along one edge (where the wire is moving to) and a brightening of the pixels at the opposite edge (where the wire has moved away, revealing the brighter sky). Simply adding scaled versions of this temporal difference results in intensity clipping artifacts for large magnification factors, such as the 75x magnification the authors apply to a video of a construction

crane (which we would rightly assume to be quite rigid) swaying imperceptibly in the wind. Mathematically speaking, the phenomenon is due to the breakdown of a Taylor Series approximation of the signal for larger motions.

The solution to this dilemma, as detailed in the following paper, is to think about amplifying the various *phases* inherent in a multi-scale decomposition of the image. Each phase difference at a given frequency band, which is due to the small motion, can be independently amplified and added back into the original signal. The authors demonstrate that this results in a perfect shift for pure sinusoids.

For a multi-scale decomposition, which groups adjacent frequencies into related sub-bands, the approximation of a shift through the addition of phase-shifted signals results in much better results than the simpler linear (all-scale) difference amplification.

While this analysis is valid for amplifying the motion seen in a single pair of video frames, improved results can be obtained by combining this analysis with selective *temporal* filtering to only amplify particular vibration frequencies. The video signal is decomposed into “three-dimensional” spatio-temporal bands, and only those

bands corresponding to the particular phenomenon of interest (vibration, swaying, breathing, and so on) are amplified, which both highlights the motions being studied and drastically reduces the amplification of video noise.

The resulting spatiotemporal motion magnification algorithms can be applied to a wide range of phenomena, including blood flow and breathing, the small motions of rigid man-made structures, and even biological (inner ear) membrane vibration.

The most surprising real-world result, however, is probably the ability to recover simple audio signals (musical notes or human speech) from the visual vibrations of a plant or bag placed in the same room as the audio source. The authors call this setup the *visual microphone*. While this may sound similar to the kinds of optical microphones used to recover sound from vibrations of windowpanes, these latter approaches use optical interferometry, while the visual microphone processes regular videos. A related approach can also be used to measure physical properties of other materials such as fabrics. Details on this and many of the other techniques discussed in the paper are provided in the ample citations.

Overall, Eulerian Motion Magnification and Analysis is a delightful tour through one of the most surprising and useful developments in computational videography in the last decade. The ability to both magnify and quantify subtle visual motions from video sequences is both a testament to the mathematical sophistication of today's multi-scale video processing algorithms and to the tremendous potential of computational photography to bring us a deeper and richer understanding of real-world phenomena. 

The following paper is a delightful tour through one of the most surprising and useful developments in computational videography in the last decade.

Richard Szeliski (szeliski@fb.com) is the director and a founding member of the Computational Photography group at Facebook, Seattle, WA.

Copyright held by author.

Eulerian Video Magnification and Analysis

By Neal Wadhwa, Hao-Yu Wu, Abe Davis, Michael Rubinstein, Eugene Shih, Gautham J. Mysore, Justin G. Chen, Oral Buyukozturk, John V. Guttag, William T. Freeman, and Frédo Durand

Abstract

The world is filled with important, but visually subtle signals. A person's pulse, the breathing of an infant, the sag and sway of a bridge—these all create visual patterns, which are too difficult to see with the naked eye. We present Eulerian Video Magnification, a computational technique for visualizing subtle color and motion variations in ordinary videos by making the variations larger. It is a *microscope for small changes* that are hard or impossible for us to see by ourselves. In addition, these small changes can be quantitatively analyzed and used to recover sounds from vibrations in distant objects, characterize material properties, and remotely measure a person's pulse.

1. INTRODUCTION

A traditional microscope takes a slide with details too small to see and optically magnifies it to reveal a rich world of bacteria, cells, crystals, and materials. We believe there is another invisible world to be visualized: that of tiny motions and small color changes. Blood flowing through one's face makes it imperceptibly redder (Figure 1a), the wind can cause structures such as cranes to sway a small amount (Figure 1b), and the subtle pattern of a baby's breathing can be too small to see. The world is full of such tiny, yet meaningful, temporal variations. We have developed tools to visualize these temporal variations in position or color, resulting in what we call a motion, or color, microscope. These new microscopes rely on computation, rather than optics, to amplify minuscule motions and color variations in ordinary and high-speed videos. The visualization of these tiny changes has led to applications in biology, structural analysis, and mechanical engineering, and may lead to applications in health care and other fields.

We process videos that may look static to the viewer, and output modified videos where motion or color changes have been magnified to become visible. In the input videos, objects may move by only 1/100th of a pixel, while in the magnified versions, motions can be amplified to span many pixels. We can also quantitatively analyze these subtle signals to enable other applications, such as extracting a person's heart rate from video, or reconstructing sound from a distance by measuring the vibrations of an object in a high-speed video (Figure 1c).

The algorithms that make this work possible are simple, efficient, and robust. Through the processing of local color or phase changes, we can isolate and amplify signals of interest. This is in contrast with earlier work to amplify small motions¹³ by computing per-pixel motion vectors and then displacing pixel values by magnified motion

vectors. That technique yielded good results but it was computationally expensive, and errors in the motion analysis would generate artifacts in the motion magnified output. As we will show, the secret to the simpler processing described in this article lies in the properties of the small motions themselves.

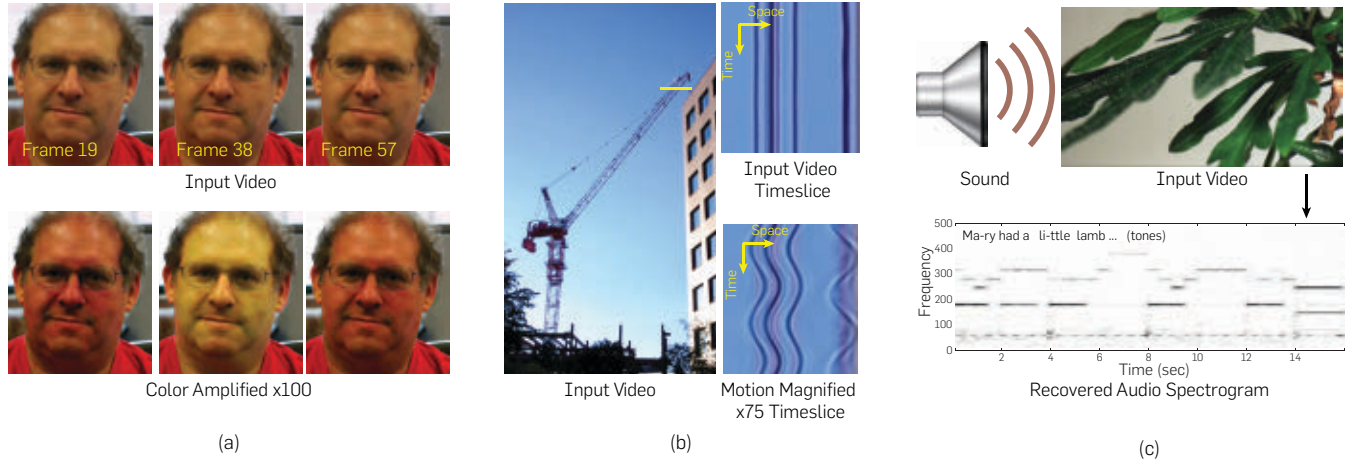
To compare our new work to the previous motion-vector work, we borrow terminology from fluid mechanics. In a *Lagrangian* perspective, the motion of fluid particles is tracked over time from the reference frame of the particles themselves, similar to observing a river flow from the moving perspective of a boat. This is the approach taken by the earlier work, tracking points in the scene and advecting pixel colors across the frame. In contrast, an *Eulerian* perspective uses a fixed reference frame and characterizes fluid properties over time at each fixed location, akin to an observer watching the water from a bridge. The new techniques we describe follow this approach by looking at temporal signals at fixed image locations.

The most basic version of our processing looks at intensity variations over time at each pixel and amplifies them. This simple processing reveals both subtle color variations *and* small motions because, for small sub-pixel motions or large structures, motion is linearly related to intensity change through a first-order Taylor series expansion (Section 2). This approach to motion magnification breaks down when the amplification factor is large and the Taylor approximation is no longer accurate. Thus, for most motion magnification applications we develop a different approach, transforming the image into a complex steerable pyramid, in which position is explicitly represented by the phase of spatially localized sinusoids. We exaggerate the phase variations observed over time, modifying the coefficients of the pyramid representation. Then, the pyramid representation is collapsed to produce the frames of a new video sequence that shows amplified versions of the small motions (Section 3). Both Eulerian approaches lead to faster processing and fewer artifacts than the previous Lagrangian approach. However, the Eulerian approaches only work well for small motions, not arbitrary ones.

Making small color changes and motions visible adds a dimension to the analysis that goes beyond simply

This Research Highlight is a high-level overview of three papers about tiny changes in videos: Eulerian Video Magnification for Revealing Subtle Changes in the World,²⁴ Phase-Based Video Motion Processing,²² and The Visual Microphone: Passive Recovery of Sound from Video.⁷

Figure 1. Apparently still videos of a face, a construction crane, and a houseplant have subtle changes that can be revealed using Eulerian video magnification and analysis. Blood flow in a man's face is revealed when the color changes are amplified (a). The construction crane's motions are revealed when amplified 75× (b). A houseplant subtly vibrates in tune with a loudspeaker playing "Mary had a little lamb." The audio is recovered from a silent video of the house plant (c).



measuring color and position changes. The visualization lets a viewer interpret the small changes, and find patterns that simply measuring numbers would not reveal. It builds intuition and understanding of the motions and changes being revealed. We show results of Eulerian video magnification in a wide variety of fields, from medicine and civil engineering to analyzing subtle vibrations due to sound. Videos and all of our results are available on our project webpage (<http://people.csail.mit.edu/mrub/vidmag/>).

2. LINEAR VIDEO MAGNIFICATION

The core idea of Eulerian video magnification is to independently process the time series of color values at each pixel. We do this by applying standard 1D temporal signal processing to each time series to amplify a band of interesting temporal frequencies, for example, around 1 Hz (60 beats per minute) for color changes and motions related to heart-rate. The new resulting time series at each pixel yield an output video where tiny changes that were impossible to see in the input, such as the reddening of a person's face with each heart beat or the subtle breathing motion of a baby, are magnified and become clearly visible.

The idea of applying temporal signal processing to each pixels' color values is a straightforward idea, and has been explored in the past for regular videos.^{10, 16} However, the results have been limited because such processing cannot handle general spatial phenomena such as large motions that involve complicated space-time behavior across pixels. When a large motion occurs, color information travels across many pixels and a Lagrangian perspective, in which motion vectors are computed, is required. One critical contribution of our work is the demonstration that in the special case of small motions, Eulerian processing can faithfully approximate their amplification. Because the motions involved are small, we can

make first-order Taylor arguments to show that linear, per-pixel amplification of color variations closely approximates a larger version of the motion. We now formalize this for the special case of 1D translational motion of a diffuse object under constant lighting, but the argument applies to arbitrary phenomena such as 3D motion and shiny objects, as we discuss below.

2.1. 1D translation

Consider a translating 1D image with intensity denoted by $I(x, t)$ at position x and time t . Because it is translating, we can express the image's intensities with a displacement function $\delta(t)$, such that $I(x, t) = f(x - \delta(t))$ and $I(x, 0) = f(x)$. Figure 2 shows the image at time 0 in black and at a later time translated to the right in blue. The goal of motion magnification is to synthesize the signal

$$\hat{I}(x, t) = f(x - (1 + \alpha)\delta(t)) \quad (1)$$

for some amplification factor α .

We are interested in the time series of color changes at each pixel:

$$B(x, t) := I(x, t) - I(x, 0). \quad (2)$$

Under the assumption that the displacement $\delta(t)$ is small, we can approximate the first term with a first-order Taylor series expansion about x , as

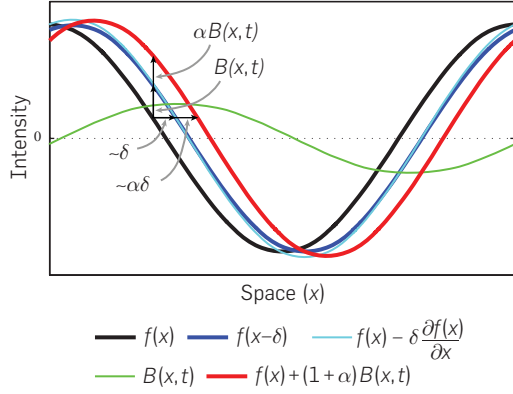
$$I(x, t) \approx f(x) - \delta(t) \frac{\partial f(x)}{\partial x}. \quad (3)$$

Because $I(x, 0) = f(x)$, the color changes at x are

$$B(x, t) \approx -\delta(t) \frac{\partial f(x)}{\partial x}. \quad (4)$$

This is the first order approximation to the well known brightness constancy equation in optical flow^{12, 14}: the intensity variation at a pixel x is the negative of the product between the displacement and the spatial gradient. This

Figure 2. Amplifying per-pixel intensity variations can approximate spatial translation. The input signal is shown at two times: $I(x, 0) = f(x)$ at time 0 (black) and $I(x, t) = f(x - \delta)$ at time t (blue). The first-order Taylor series expansion of $I(x, t)$ around x approximates the translated signal (cyan). The pointwise difference between the frames (green) is amplified and added to the original signal to generate a larger translation (red). Here, the amplification factor α is 1, amplifying the motion by 100%.



can be seen as a right triangle in Figure 2, whose legs are the temporal intensity variation (vertical edge marked $B(x, t)$) and the displacement (horizontal edge marked δ) and whose hypotenuse (blue curve between the legs) has slope equal to the image's spatial derivative $\left(\frac{\partial f(x)}{\partial x}\right)$.

In our processing, we amplify the color change signal $B(x, t)$ by α and add it back to $I(x, t)$, resulting in the processed signal (red in Figure 2):

$$\tilde{I}(x, t) = I(x, t) + \alpha B(x, t). \quad (5)$$

Combining equations (3)–(5), we have

$$\tilde{I}(x, t) \approx f(x) - (1 + \alpha)\delta(t) \frac{\partial f(x)}{\partial x}. \quad (6)$$

As long as $(1 + \alpha)\delta$ is small enough that a first-order Taylor expansion is valid, we can relate the previous equation to motion magnification (Eq. 1). It is simply

$$\tilde{I}(x, t) \approx f(x - (1 + \alpha)\delta(t)). \quad (7)$$

This shows that this processing magnifies motions. The spatial displacement $\delta(t)$ between frames of the video at times 0 and t , has been amplified by a factor of $(1 + \alpha)$.

2.2. General case

Consider a subtle, temporal phenomenon, for example, 3D translation, rotation, or the motion of light, parameterized by a vector θ (perhaps representing the position or orientation of objects or lights) that evolves over time as $\theta(t)$. These parameters can be mapped to image intensities $I(x, t)$ via a function $f(x, \theta(t))$ for all spatial locations x . If f is a differentiable function of the parameters θ and the changes in the parameters are small, then the video I can be approximated by its first order Taylor expansion around $\theta(0)$

$$I(x, t) \approx f(x, \theta(0)) + \nabla f(x, \theta(0))^T (\theta(t) - \theta(0)). \quad (8)$$

That is, each pixel in the video signal is linearly related to the deviation of the parameters θ from their initial value. If we amplify by α the difference between the image at time t and at time 0, we get

$$\tilde{I}(x, t) := f(x, \theta(0)) + (1 + \alpha) \nabla f(x, \theta(0))^T (\theta(t) - \theta(0)). \quad (9)$$

By the same analysis as before, this is approximately equal to a new video in which the variations in θ are larger by a factor $1 + \alpha$. This shows that linear Eulerian video magnification can be used to magnify many subtle, temporal phenomena. It is agnostic to the underlying imaging model and can even work in cases where brightness constancy is not true as long as the changes are small.

2.3. Limitations of the linear approach

Linear amplification relies on a first-order Taylor expansion, which breaks down when either the amplification factor or the input motion is too large. For overly large amplification factors, the magnified video overshoots and undershoots the video's white and black levels causing clipping artifacts near edges where the second derivative $\left(\frac{\partial^2 f(x)}{\partial x^2}\right)$ is non-negligible (Figure 6a). When the input motion is too large, the initial Taylor expansion is inaccurate (Eq. 3) and the output contains ghosting artifacts instead of magnified motions.

A second limitation is that noise in the video is amplified. For example, suppose the intensity value $I(x, t)$ has an independent white Gaussian noise term $n(x, t)$ of variance σ^2 added to it. The difference between the frame at time t and at time 0 then contains the noise term, $n(x, t) - n(x, 0)$, with noise variance $2\sigma^2$. This noise term gets amplified by a factor α and the output video has noise of variance $2\alpha^2\sigma^2$, a much larger amount than in the input video (Figure 7b).

In Wu et al.,²⁴ noise amplification was partially mitigated by reducing the amplification of high spatial-frequency temporal variations, assuming that they are mostly noise rather than signal. This is done by constructing a Laplacian pyramid of the temporal variations and using a lower amplification factor for high spatial-frequency levels. Spatially lowpassing the temporal variations produces comparable results. A thorough noise analysis of this approach is available in the appendix of Wu et al.²⁴ and more information about signal-to-noise ratios is given here in Section 4.

3. PHASE-BASED MAGNIFICATION

The appeal of the Eulerian approach to video magnification is that it independently processes the time series of color values at each pixel and does not need to explicitly compute motions. However, its reliance on first-order approximations limits its scope, and its use of linear amplification increases noise power. In this section, we seek to continue using the Eulerian perspective of motion analysis—processing independent time series at fixed reference locations. But, we want to do so in a representation that better handles motions and is less prone to noise.

In the case of videos that are global translations of a frame over time, there is a representation, that is, exactly what we want: the Fourier series. Its basis functions are complex-valued sinusoids that, by the Fourier shift theorem, can be translated exactly by shifting their phase (Figure 3a,c). However, using the Fourier basis would limit us to only being able to handle the same translation across the entire frame, precluding the amplification of complicated spatially-varying motions. To handle such motions, we instead use spatially-local complex sinusoids implemented by a wavelet-like representation called the complex steerable pyramid.^{19, 20} This representation decomposes images into a sum of complex wavelets corresponding to different scales, orientations, and positions. Each wavelet has a notion of local amplitude and local phase, similar to the amplitude and phase of a complex sinusoid (Figure 4a). The key to our new approach is to perform the same 1D temporal signal processing and amplification described earlier on the local phase of each wavelet, which directly corresponds to local motion as we discuss below.

3.1. Simplified global case

To provide intuition for how phase can be used to magnify motion, we work through a simplified example in which a global 1D translation of an image is magnified using the phase of Fourier basis coefficients (Figure 3).

Let image intensity $I(x, t)$ be given by $f(x - \delta(t))$ where $\delta(0) = 0$. We decompose the profile $f(x)$ into a sum of complex coefficients times sinusoids using the Fourier transform

$$f(x) = \sum_{\omega} A_{\omega} e^{i\phi_{\omega}} e^{-i\omega x}. \quad (10)$$

Because the frames of I are translations of f , their Fourier transform is given by a phase shift by $\omega\delta(t)$:

$$I(x, t) = \sum_{\omega} A_{\omega} e^{i\phi_{\omega}} e^{-i\omega(x-\delta(t))} = \sum_{\omega} A_{\omega} e^{i(\phi_{\omega} + \omega\delta(t))} e^{-i\omega x}, \quad (11)$$

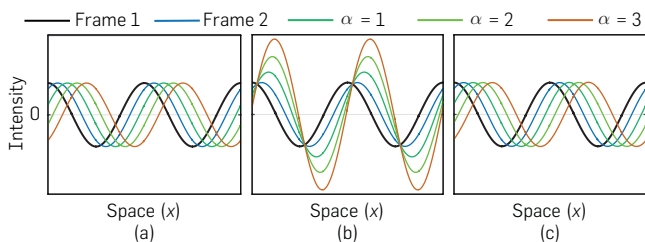
where the phase of these coefficients becomes $\phi_{\omega} + \omega\delta(t)$. If we subtract the phase at time 0 from the phase at time t , we get the phase difference

$$\omega\delta(t), \quad (12)$$

which is proportional to the translation. Amplifying this phase difference by a factor α and using it to shift the Fourier coefficients of $I(x, t)$ yields

$$\sum_{\omega} A_{\omega} e^{i\phi_{\omega} + (1+\alpha)\omega\delta(t)} e^{-i\omega x} = f(x - (1+\alpha)\delta(t)), \quad (13)$$

Figure 3. Phase-based motion magnification is perfect for Fourier basis functions (sinusoids). In these plots, the initial displacement is $\delta(t) = 1$. (a) True Amplification. (b) Linear. (c) Phase-Based.



a new image sequence in which the translations have been *exactly* magnified.

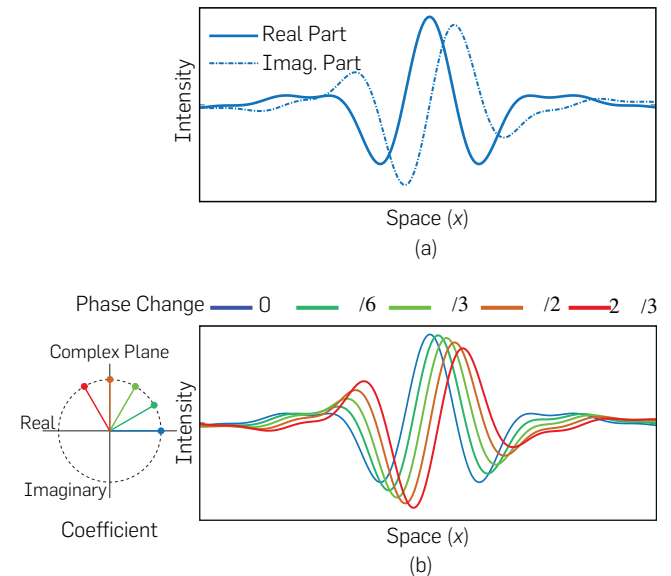
Phase-based magnification works perfectly in this case because the motions are global and because the transform breaks the image into a representation consisting of exact sinusoids (formally, the Fourier transform diagonalizes the translation operator). In most cases, however, the motions are not global, but local. This is why we break the image into local sinusoids using the complex steerable pyramid.

3.2. Complex steerable pyramid

The complex steerable pyramid^{19, 20} is a complex, over-complete linear transform. It decomposes an image into a set of coefficients corresponding to basis functions that are simultaneously localized in position, spatial scale and orientation. The image is reconstructed by multiplying the coefficients by the basis functions and summing the real parts.

The transform is best-described by its self-similar basis functions. Each one is a translation, dilation, or rotation of another. So, it is sufficient to look at just one, a 1D version of which is shown in Figure 4a. It resembles an oriented complex sinusoid windowed by a Gaussian envelope. The complex sinusoid provides locality in frequency while the windowing provides locality in space. Each basis function is complex, consisting of a real, even-symmetric part (cosine) and an imaginary, odd-symmetric part (sine). This gives rise to a notion of local amplitude and local phase as opposed to the global amplitude and phase of Fourier basis functions. We use only a half-circle of orientations because basis functions at antipodal orientations ($\theta, \theta + \pi$) yield redundant, conjugate coefficients.

Figure 4. Increasing the phase of complex steerable pyramid coefficients results in approximate local motion of the basis functions. A complex steerable pyramid basis function (a) is multiplied by several complex coefficients of constant amplitude and increasing phase to produce the real part of a new basis function, that is, approximately translating (b).



3.3. Local phase shift is local translation

The link between local phase shift and local translation has been studied before in papers about phase-based optical flow.^{9,11} Here, we demonstrate how local phase shift approximates local translation for a single basis function in a manner similar to the global phase-shift theorem of Fourier bases. We model a basis function as a Gaussian window multiplied by a complex sinusoid

$$e^{\frac{-x^2}{(2\sigma^2)}} e^{-i\omega x}, \quad (14)$$

where σ is the standard deviation of the Gaussian envelope and ω is the frequency of the complex sinusoid. In the complex steerable pyramid, the ratio between σ and ω is fixed because the basis functions are self-similar. Low frequency wavelets have larger windows.

Changing the *phase* of the basis element by multiplying it by a complex coefficient $e^{i\phi}$ results in

$$e^{\frac{-x^2}{(2\sigma^2)}} e^{-i\omega x} \times e^{i\phi} = e^{\frac{-x^2}{(2\sigma^2)}} e^{-i\omega(x-\phi/\omega)}. \quad (15)$$

The complex sinusoid under the window is translated, which is approximately a translation of the whole basis function by $\frac{\phi}{\omega}$ (Figure 4b).

Conversely, the phase difference between two translated basis elements is proportional to translation. Specifically, suppose we have a basis element and its translation by δ :

$$e^{\frac{-x^2}{(2\sigma^2)}} e^{-i\omega x}, e^{\frac{-(x-\delta)^2}{(2\sigma^2)}} e^{-i\omega(x-\delta)}. \quad (16)$$

The local phase of each element only depends on the argument to the complex exponential, and is $-\omega x$ in the first case and $-\omega(x-\delta)$ in the second. The phase difference is then $\omega\delta$, which is directly proportional to the translation. Local phase shift can be used both to analyze tiny translations and synthesize larger ones.

3.4. Our method

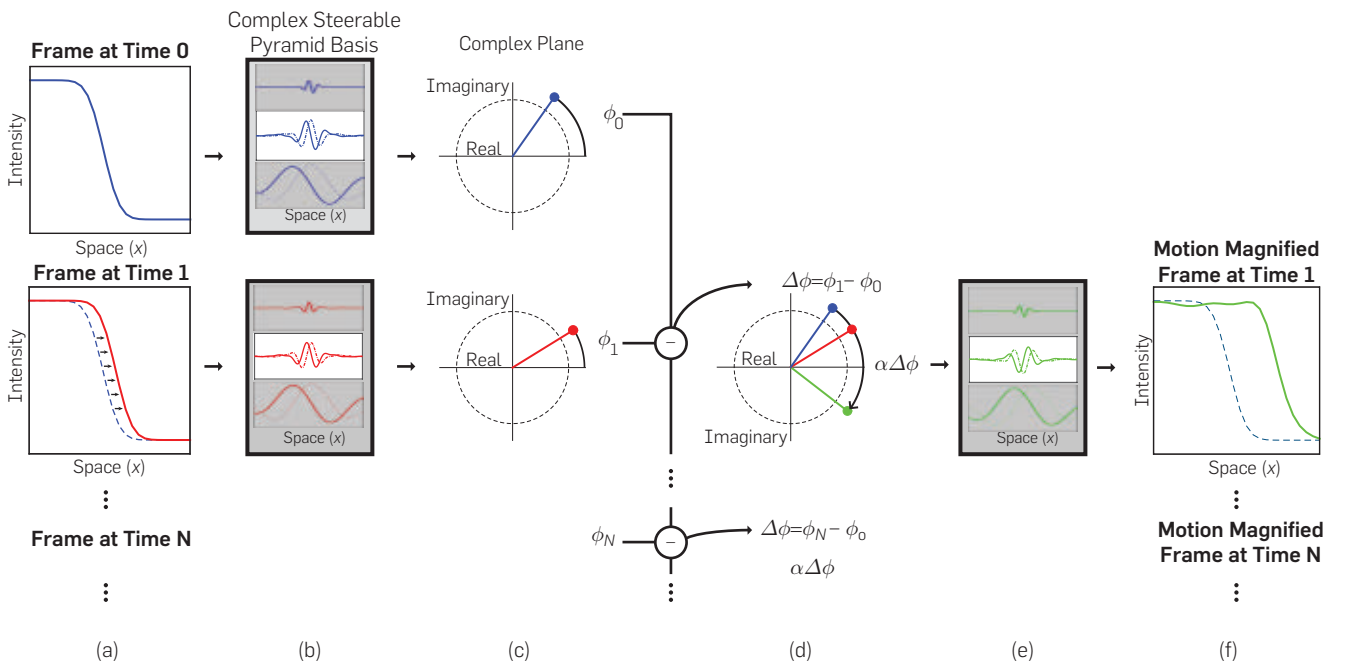
The observation that local phase differences can be used to manipulate local motions motivates our pipeline. We take an image sequence, project each frame onto the complex steerable pyramid basis and then independently amplify the phase differences between *all* corresponding basis elements. This is identical to the linear amplification pipeline except that we have changed the representation from intensities to local spatial phases.

To illustrate the pipeline, consider again an image sequence $I(x, t)$, in which the frame at time 0 is $f(x)$ and the frames at time t are translations $f(x - \delta(t))$ (Figure 5a). In our first step, we project each frame onto the complex steerable pyramid basis (Figure 5b), which results in a complex coefficient for every scale r , orientation θ and spatial location x, y , and time t . Because the coefficients are complex, they can be expressed in terms of amplitude $A_{r,\theta}$ and phase $\phi_{r,\theta}$ as

$$A_{r,\theta}(x, y, t) e^{i\phi_{r,\theta}(x, y, t)}. \quad (17)$$

In Figure 5c, we show coefficients at a specific location, scale, and orientation in the complex plane at times 0 and 1.

Figure 5. A 1D example illustrating how the local phase of complex steerable pyramid coefficients is used to amplify the motion of a subtly translating step edge. Frames (two shown) from the video (a) are transformed to the complex steerable pyramid representation by projecting onto its basis functions (b), shown in several spatial scales. The phases of the resulting complex coefficients are computed (c) and the phase differences between corresponding coefficients are amplified (d). Only a coefficient corresponding to a single location and scale is shown; this processing is done to all coefficients. The new coefficients are used to shift the basis functions (e) and a reconstructed video is produced in which the motion of the step edge is magnified (f).



Because the two frames are slight translations of each other, each coefficient has a slight phase difference. This is illustrated in Figure 5c, in which the coefficients have roughly the same amplitude but different phases. The next step of our processing is to take the phase differences between the coefficients in the video and those of a reference frame, in this case the frame at time 0:

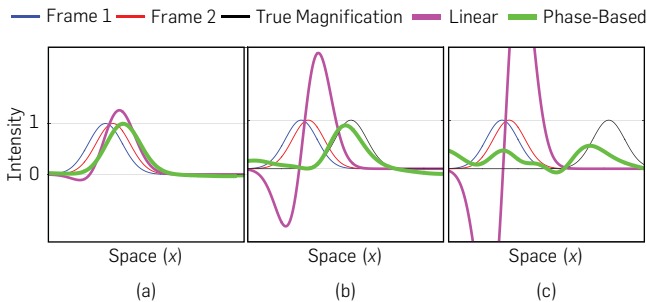
$$\Delta\phi_{r,\theta}(x, y, t) = \phi_{r,\theta}(x, y, t) - \phi_{r,\theta}(x, y, 0). \quad (18)$$

These phase differences are amplified by a factor α (Figure 5d), which yields a new set of coefficients for each frame, in which the amplitudes are the same, but the phase differences from the reference frame are larger. We reconstruct the new frames using these coefficients by multiplying them by the basis functions (Figure 5e). Then, we sum the real part to get new frames, in which the translations—and therefore the motions in the video—are amplified (Figure 5f).

Amplifying phase differences rather than pixel intensity differences has two main advantages: (a) it can support larger amplification factors, and (b) noise amplitude does not get amplified. In Figure 6, we show the two different methods being used to amplify the motions of a 1D Gaussian bump. Both methods work well for small amplification factors (Figure 6a). For larger amplification factors, amplifying raw pixel differences results in the signal overshooting the white level and undershooting the black level resulting in intensity clipping. In contrast, amplifying phase differences allows us to push the Gaussian bump much farther (Figure 6b). At very high amplification levels, the different spatial scales of the bump break apart because the high frequency components cannot be pushed as far as the lower frequency components (Figure 6c).

In Figure 7, we show the effect of both methods on a video, which consists of independent and identically distributed (iid) Gaussian noise. Unlike the linear method which increases noise power, the phase based method preserves noise power preventing objectionable artifacts in the motion magnified output. For these reasons, we found that amplifying phase differences rather than pixel differences is a better approach for magnifying small motions.

Figure 6. For non-periodic structures, both methods work for small amplification, $\alpha = 1.5$ (a). The phase-based method supports amplification factors four times as high as the linear method and does not suffer from intensity clipping artifacts, $\alpha = 6$ (b). For large amplification, different frequency bands break up because the higher frequency bands have smaller windows, $\alpha = 14$ (c).



3.5. Riesz pyramids

Using phase in the complex steerable pyramid to motion magnify videos can be slow because the representation is much larger than the input. We have developed another method that is similar in spirit and produces videos of almost the same quality. However, it is much faster and is capable of running in real-time on a laptop. More details about this can be found in this paper²³ on Riesz Pyramids.

4. AMPLIFYING THE RIGHT SIGNAL

Maximizing the signal-to-noise ratio of the temporal variations we amplify, whether local phase changes or color changes, is the key to good performance. We improve SNR by temporally and spatially filtering the variations to remove components that correspond to noise and keep those that correspond to signal. The temporal filtering also gives a way to isolate a signal of interest as different motions often occur at different temporal frequencies. A baby's squirming might be at a lower temporal frequency than her breathing.

Temporal narrowband linear filters provide a good way to improve signal-to-noise ratios for motions that occur in a narrow range of frequencies, such as respiration and vibrations. To prevent phase-wrapping issues when using these filters, we first unwrap the phases in time. The filters can also be used to isolate motions in an object that correspond to different frequencies. For example, a pipe vibrates at a preferred set of modal frequencies, each of which has a different spatial pattern of vibration. We can use video magnification to reveal these *spatial* patterns by amplifying the motions only corresponding to a range of *temporal* frequencies. A single frame from each motion magnified video is shown in Figure 8, along with the theoretically expected shape.²¹

Spatially smoothing the motion signal often improves signal-to-noise ratios. Objects tend to move coherently in local image patches and any deviation from this is likely noise. Because the phase signal is more reliable when the amplitude of the complex steerable pyramid coefficients is higher, we perform an amplitude-weighted Gaussian blur:

$$\frac{((\Delta\phi)A) \star K_\rho}{A \star K_\rho} \quad (19)$$

where K_ρ is a Gaussian convolution kernel given by $\exp\left(-\frac{x^2 + y^2}{2\rho^2}\right)$. The indices of A and ϕ have been suppressed for readability.

Figure 7. Comparison between linear and phase-based Eulerian motion magnification in handling noise. (a) A frame in a sequence of iid noise. In both (b) and (c), the motion is amplified by a factor of 50, where (b) amplifies changes linearly, while (c) uses the phase-based approach.

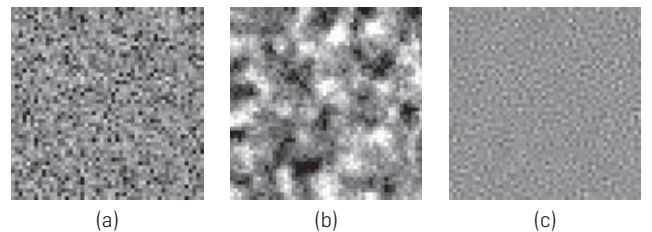
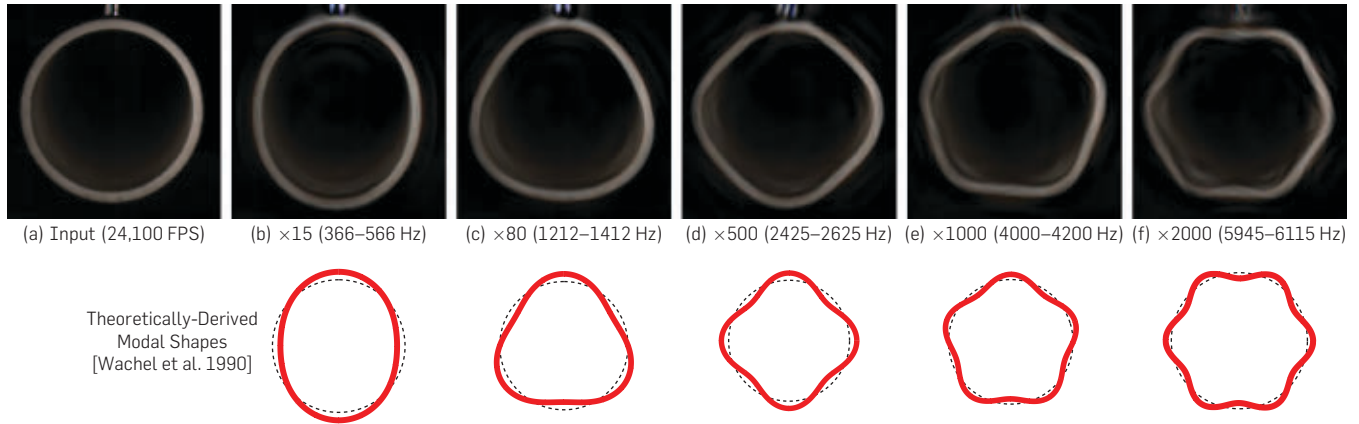


Figure 8. Isolating different types of *spatial* motions with *temporal* filtering. We took a high-speed video of a pipe being struck with a hammer. A frame from this video is shown in (a). The motions at several frequencies were magnified to isolate different modal shapes of the pipe. In (b)–(f), a frame from each of the motion magnified videos is shown. Below, the theoretically-derived modal shapes are shown in red overlaid, for comparison, over a perfect circle in dotted black.



We applied this processing to all of our motion magnification videos with ρ equal to 2 pixels in each pyramid level.

Because oversmoothing can shape even white noise into a plausible motion signal, it becomes important to verify whether the motions we are amplifying are indeed real. We have done many experiments comparing the visual motion signal with the signal recorded by accurate point-measurement devices, such as an accelerometer or laser vibrometer and the signals are always in agreement, validating that the motions are real.^{2–4, 22} In addition, there are many videos where the motion is spatially coherent at a scale beyond that imposed by spatial smoothing (e.g., the pipes in Figure 8). This is unlikely to happen by chance and provides further evidence for the correctness of the amplified videos.

We can only recover motions that occur at frequencies less than the temporal Nyquist frequency of the camera. If the motions are too fast, only an aliased version of them gets amplified. In the special case that the motions occur at a single temporal frequency, aliasing can be useful. It makes such motions appear slower, which permits the visualization of fast vibrations in real-time, for example, the resonance of a wine glass.²³ However, in general we cannot recover a meaningful signal if the frames are temporally undersampled.

5. A BIG WORLD OF SMALL CHANGES

The world is full of subtle changes that are invisible to the naked eye. Video magnification allows us to reveal these changes by magnifying them. We present a selection of our magnification results and extensions of our techniques by us and other authors below.

As the heart beats, blood flows in and out of the face, changing its color slightly. This color change is very subtle, typically only half a gray-level. However, the human pulse occurs in a narrow band of temporal frequencies and is spatially smooth. For the man in Figure 1a, we can isolate signal from noise by temporally filtering the color variations in a passband of 50–60 beats per minute (0.83–1 Hz) and then spatially lowpassing them. Amplifying the result by $100\times$

produces a color-amplified video, in which the human pulse is visible (Figure 1a). In addition to visualizing the pulse, we can plot the filtered color changes at a point on the man's forehead to get a quantitative measurement of pulse (Figure 3.8 in Ref.¹⁷). The beating of the human heart also produces subtle motions throughout the body. We were able to visualize the pulsing of the radial and ulnar arteries in a video of a wrist (Figure 7 in Ref.²⁴). Amir-Khalili et al. and McLeod et al. have also quantitatively analyzed subtle color and motion changes using methods inspired by the ones proposed here to identify faintly pulsing blood vessels in surgical videos,^{1, 15} which may be clinically useful.

Our methods can also be used to reveal the invisible swaying of structures. We demonstrate this in a video of a crane taken by an ordinary DSLR camera (Figure 1b). The crane looks completely still in the input video. However, when the low-frequency (0.2–0.4 Hz) motions are magnified $75\times$, the swaying of the crane's mast and hook become apparent. In an extension to the work described here, Chen et al. quantitatively analyze structural motions in videos to non-destructively test their safety.³ They do this by recovering modal shapes and frequencies of structures (similar to Figure 8) based on local phase changes in videos and use these as markers for structural damage.

Video magnification has also contributed to new scientific discoveries in biology. Sellon et al. magnified the subtle motions of an in-vitro mammalian tectorial membrane,¹⁸ a thin structure in the inner ear. This helped explain this membrane's role in frequency selectivity during hearing.

6. THE VISUAL MICROPHONE

One interesting source of small motions is sound. When sound hits an object, it causes that object to vibrate. These vibrations are normally too subtle and too fast to be seen, but we can sometimes reveal them in motion magnified, high-speed videos of the object (Figure 9a). This shows that sound can produce a *visual* motion signal. Video magnification gives us a way to visualize this signal, but we can also

quantitatively analyze it to recover sound from silent videos of the objects (Figure 9b). For example, we can recover intelligible speech and music from high-speed videos of a vibrating potato chip bag or houseplant (Figures 1 and 10). We call this technique The Visual Microphone.

Figure 9. Revealing sound-induced vibrations using motion magnification and recovering audio using the visual microphone. A pure-tone version of “Mary Had a Little Lamb” is played at a chip bag and the motions corresponding to each note are magnified separately. Time slices of the resulting videos, in which the vertical dimension is space and the horizontal dimension is time, are used to produce a visual spectrogram (a), which closely matches the spectrogram of the recovered audio (b).

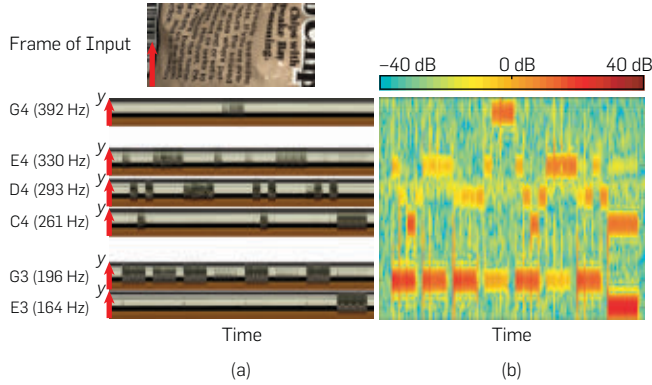
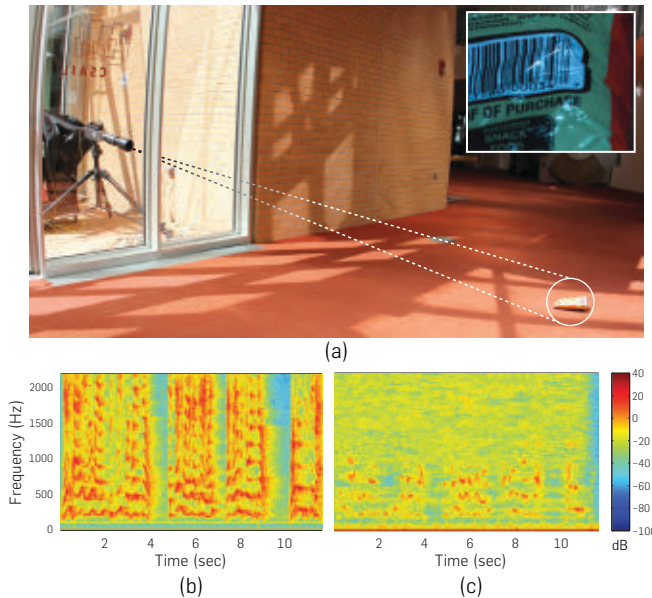


Figure 10. Speech recovered from a 4kHz silent video of a bag of chips filmed through soundproof glass. The chip bag (on the bottom right in (a)) is lit by natural sunlight. The camera (on the left in (a)) is outside behind sound-proof glass. A frame from the recorded video (400 × 480 pixels) is shown in the inset. The speech “Mary had a little lamb...Welcome to SIGGRAPH!” was spoken by a person near the bag of chips. (b) The spectrogram of the source sound recorded by a standard microphone near the chip bag and (c) the spectrogram of our recovered sound. The recovered sound is noisy but comprehensible (audio clips are available on the visual microphone project webpage).



A challenge is that the vibrations are incredibly small, on the order of micrometers for 80 dB sound, and noise in the video can easily overwhelm this signal. Narrowband temporal filtering only works for narrowband sounds, not general ones that contain all frequencies. However, we want to recover a 1D audio signal, not magnify the spatial pattern of the motions. This means we can spatially combine information across the *entire* frame to attain intelligible SNR. This works because at most audible frequencies, the wavelength of the sound (10 cm–1.2 m for telephone-quality audio) is much larger than the visible portion of the object in a video. For lightweight objects, the motion caused by sound is largely coherent across each frame of the video.

Based on this observation, we seek to recover a global motion signal $R(t)$ of the object. We measure local motions by using local phase variations. We project each frame of the input video on to the complex steerable pyramid basis and then compute the time series of local phase variations $\Delta\phi_{r,\theta}(x, y, t)$ for all spatial scales r , orientations θ and positions x, y .

To place higher confidence in local motions that come from regions with localizable features, we weigh the local phase variations by the square of the amplitudes of the corresponding coefficients. We then take the average of these weighted local signals over the dimensions of the complex steerable pyramid:

$$R(t) = \sum_{r,\theta,x,y} A_{r,\theta}(x, y, t)^2 \Delta\phi_{r,\theta}(x, y, t). \quad (20)$$

This signal measures the average motion of the object and is related to sound via an object-specific transfer function.⁷ This signal can be played as intelligible sound and can be further improved with standard audio denoising techniques.

We conducted a variety of experiments to explore when, and how well, this method was able to recover sound. In each experiment, an object was filmed with a high-speed camera while being exposed to sound from either a nearby loudspeaker or a person’s voice. Experiments were calibrated with a decibel meter and evaluated using perception-based metrics from the audio processing literature. We found that lightweight objects, which move readily with sound (e.g., a bag of chips, the leaves of a plant) yielded the best results. Heavier objects (e.g., bricks, an optical bench) produced much weaker results, suggesting that unintended motion, like camera shake, was not a significant factor. In our most ambitious experiment, we were able to recover human speech from a bag of chips 3–4 m away (Figure 10). More information about our experiments are in Ref.⁷

In one experiment, we played a pure-tone version (synthesized from MIDI) of “Mary had a Little Lamb” at a chip bag. Because the sound contained only pure-tones, we were able to motion magnify the chip bag in narrow temporal bands corresponding to the tones to produce six processed videos that together form a *visual spectrogram*. We show slices in time of the motion magnified videos (Figure 9a) and display them next to the recovered sound’s spectrogram (Figure 9b).

Vibrations of an object can also be used to learn about its physical properties. In follow-up work, Davis and Bouman et al.⁵ use the same method to analyze small vibrations in

objects and discern the material properties (stiffness, area weight, and elasticity) of fabrics from videos. Davis et al.⁶ also used tiny motions to learn an image-space model of how an object vibrates and used that to perform simulations of how it would behave if stimulated.

7. LIMITATIONS

The Eulerian approach to motion magnification is robust and fast, but works primarily when the motions are small. If the motions are large, this processing can introduce artifacts. However, one can detect when this happens and suppress magnification in this case.²² Elgharib et al.⁸ also demonstrate it is possible to magnify tiny motions in the presence of large ones by first stabilizing the video. There are limits to how well spatio-temporal filtering can remove noise and amplified noise can cause image structures to move incoherently.

8. CONCLUSION

Eulerian video magnification is a set of simple and robust algorithms that can reveal and analyze tiny motions. It is a new type of microscope, not made of optics, but of software taking an ordinary video as input and producing one in which the temporal changes are larger. It reveals a new world of tiny motions and color changes showing us hidden vital signs, building movements and vibrations due to sound waves. Our visualization may have applications in a variety of fields such as healthcare, biology, mechanical engineering, and civil engineering.

Acknowledgments

We thank Quanta Computer, Shell research, QCRI, NSF CGV-1111415, the DARPA SCENICC program, the NDSeg fellowship, the Microsoft Research fellowship, and Cognex for funding this research. We also thank Dirk Smit, Steve Lewin-Berlin, Guha Balakrishnan, Ce Liu, and Deqing Sun for their helpful suggestions. We thank Michael Feng at Draper Laboratory for loaning us a laser vibrometer and Dr. Donna Brezinksi, Dr. Karen McAlmon and the Winchester Hospital staff for helping us to collect videos of newborns. ■

References

1. Amir-Khalili, A., Peyrat, J.-M., Abinayed, J., Al-Alao, O., Al-Ansari, A., Hamarneh, G., Abugharbieh, R. Auto localization and segmentation of occluded vessels in robot-assisted partial nephrectomy. In *MICCAI 2014* (2014). Springer, 407–414.
2. Chen, J.G., Wadhwa, N., Cha, Y.-J., Durand, F., Freeman, W.T., Buyukozturk, O. Structural modal identification through high speed camera video: Motion magnification. In *Topics in Modal Analysis I*. Volume 7 (2014). Springer, 191–197.
3. Chen, J.G., Wadhwa, N., Cha, Y.-J., Durand, F., Freeman, W.T., Buyukozturk, O. Modal identification of simple structures with high-speed video using motion magnification. *Journal of Sound and Vibration* 345 (2015), 58–71.
4. Chen, J.G., Wadhwa, N., Durand, F., Freeman, W.T., Buyukozturk, O. Developments with motion magnification for structural modal

identification through camera video. In *Dynamics of Civil Structures*. Volume 2 (2015). Springer, 49–57.

5. Davis, A., Bouman, K.L., Chen, J.G., Rubinstein, M., Durand, F., Freeman, W.T. Visual vibrometry: Estimating material properties from small motions in video. In *CVPR 2015* (2015), 5335–5343.
6. Davis, A., Chen, J.G., Durand, F. Image-space modal bases for plausible manipulation of objects in video. *ACM Trans. Graph.* 34, 6 (2015), 239.
7. Davis, A., Rubinstein, M., Wadhwa, N., Mysore, G.J., Durand, F., Freeman, W.T. The visual microphone: Passive recovery of sound from video. *ACM Trans. Graph.* 33, 4 (2014), 79.
8. Elgharib, M.A., Hefeeda, M., Durand, F., Freeman, W.T. Video magnification in presence of large motions. In *CVPR 2015* (2015), 4119–4127.
9. Fleet, D.J., Jepson, A.D. Computation of component image velocity from

local phase information. *Int. J. Comput. Vision* 5, 1 (1990), 77–104.

10. Fuchs, M., Chen, T., Wang, O., Raskar, R., Seidel, H.-P., Lensch, H.P. Real-time temporal shaping of high-speed video streams. *Comput. Graph.* 34, 5 (2010), 575–584.
11. Gautama, T., Van Hulle, M.M. A phase-based approach to the estimation of the optical flow field using spatial filtering. *IEEE Trans. Neural Netw.* 13, 5 (2002), 1127–1136.
12. Horn, B.K., Schunck, B.G. Determining optical flow. In *1981 Technical Symposium East* (1981). International Society for Optics and Photonics, 319–331.
13. Liu, C., Torralba, A., Freeman, W.T., Durand, F., Adelson, E.H. Motion magnification. *ACM Trans. Graph.* 24, 3 (2005), 519–526.
14. Lucas, B.D., Kanade, T., et al. An iterative image registration technique with an application to stereo vision. In *IJCAI*. Volume 81 (1981), 674–679.
15. McLeod, A.J., Baxter, J.S., de Ribaupierre, S., Peters, T.M. Motion magnification for endoscopic surgery. In *SPIE Medical Imaging* (2014), 90360C–90360C.
16. Poh, M.-Z., McDuff, D.J., Picard, R.W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express* 18, 10 (2010), 10762–10774.
17. Rubinstein, M. Analysis and visualization of temporal variations in video. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA (2014).
18. Sellon, J.B., Farrahi, S., Ghaffari, R., Freeman, D.M. Longitudinal spread of mechanical excitation through tectorial membrane traveling waves. *Proc. Natl. Acad. Sci.* 112, 42 (2015), 12968–12973.
19. Simoncelli, E.P., Freeman, W.T. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *ICIP 1995* (1995), 3444.
20. Simoncelli, E.P., Freeman, W.T., Adelson, E.H., Heeger, D.J. Shifttable multi-scale transforms. *IEEE Trans. Info. Theory* 2, 38 (1992), 587–607.
21. Wachel, J., Morton, S.J., Atkins, K.E. Piping vibration analysis. In *Proceedings of the 19th Turbomachinery Symposium*. (1990).
22. Wadhwa, N., Rubinstein, M., Durand, F., Freeman, W.T. Phase-based video motion processing. *ACM Trans. Graph.* 32, 4 (2013), 80.
23. Wadhwa, N., Rubinstein, M., Durand, F., Freeman, W.T. Riesz pyramids for fast phase-based video magnification. In *ICCP 2014* (2014). IEEE, 1–10.
24. Wu, H.-Y., Rubinstein, M., Shih, E., Guttag, J. V., Durand, F., Freeman, W.T. Eulerian video magnification for revealing subtle changes in the world. *ACM Trans. Graph.* 31, 4 (2012), 65.

Neal Wadhwa, Abe Davis, John V. Guttag, William T. Freeman and Fredo Durand, MIT Computer Science and Artificial Intelligence Laboratory

Michael Rubinstein, Google Research

Eugene Shih, Cambridge Mobile Telematics

Gautham J. Mysore, Adobe Research

Justin G. Chen and Oral Buyukozturk, MIT Department of Civil and Environmental Engineering

Hao-Yu Wu, Amazon A9

Technical Perspective

Mapping the Universe

By Valentina Salapura

“IF PEOPLE SAT outside and looked at the stars each night, I’ll bet they’d live a lot differently.”

—Bill Watterson, *Calvin and Hobbes*

Who has not been in awe and had their breath taken away while gazing at the star-studded night sky? Looking into infinity helps us realize the universe is more vast than we could ever contemplate and helps put our daily lives in perspective to a much larger cosmos.

Dark energy and dark matter, the observable universe and its beginnings, the structure and formation of galaxies and stars—these are just some of the topics that computational cosmologists work on in order to understand the universe. As the universe expanded and temperatures cooled after the early hot phase, the present complex structure of the universe started to form. The expansion of the universe is now accelerating due to a mysterious dark energy. The mass in the universe is dominated by dark matter; galaxies and stars are located within extended clumps of this form of matter, whose ultimate nature is unknown.

Modern cosmological simulations of ever-higher accuracy and larger complexity have an insatiable appetite for computing resources—processors, memory, and high-performance networks. Simulating the universe will stretch the capabilities of the most advanced supercomputers for the foreseeable future. This demand is driven by the desire for increased realism and accuracy, and by the large amount of simulated and observed data that must be jointly analyzed.

The following paper describes the Hardware/Hybrid Accelerated Cosmology Code (HACC) framework. The architecture of the framework is organized around particle and grid methods. HACC uses a novel algorithmic structure to map code onto multiple supercomputer platforms, built around different node-level architectures, whether heterogeneous CPU/GPU or multi/many-core systems.

During the last decade, the manner in which computers are built has changed drastically. Hardware evolution embraced disruptive change as increasing clock frequency was no longer a viable path to increasing performance. Before, the approach to achieve increased performance was to follow Dennard scaling. Transistor sizes would be scaled down and a smaller feature size would result in higher circuit operational frequency. Computer chips would run at higher frequency, and would deliver higher performance.

The frequency increase of computer chips stopped due to power density constraints; at the same time shrinking transistor sizes also hit a wall due to the eventual constraint set by the atomic nature of matter. Thus, computer architects in pursuit of building high-performance computer systems turned toward increasing the number of cores, rather than increasing the performance of each core. In this brave new world, high performance was to be achieved through extreme parallelism rather than high frequency. The most energy-efficient approach to reaching maximum performance is to increase parallelism using efficient low-frequency processors. For example, using a large number of very simple processors with limited capabilities—accelerators—is very energy efficient. The ensuing revolution changed how computers are built. Accelerators are now common, making computing inexpensive, while memory and communication remain relatively more expensive, thereby changing the balance of memory and communication per compute unit.

Revolutionary new architectures appeared in the middle of the last decade—such as the Roadrunner supercomputer with IBM’s Cell processor chip, the first supercomputer to cross the petaflop barrier. The Cell chip introduced an asymmetric architecture, containing a general-purpose Power

core connected to a number of simple accelerators. This new architecture at the time required a different programming approach, and scientists—including members of the HACC team—started rewriting their code in order to handle architectural diversity.

After Roadrunner, the team ran their code on two very different machines: on the BlueGene/Q—the third generation of the IBM Blue Gene supercomputer running on more than 1.5 million compute cores, and on Cray’s Titan, a hybrid Cray XK7 system with GPU accelerators. HACC demonstrated high performance and good scaling on these two different architectures.

The evolution of hardware architecture poses a number of challenges for software developers, and particularly for scientific code developers. These are typically small communities that maintain their codes over many years. As access to supercomputers is typically limited, granted time on these machines needs to be spent wisely, by running performance-optimized codes. This puts a requirement on domain scientists to adapt and optimize their code for the target machine.

In order to adapt their codes to new machines, the scientists must understand all levels of system architecture of the target machine. The authors explore this topic. How does one code physical models for the vast variety of supercomputers, for very different architectures—architectures with or without accelerators, and with very different ratios of computing/memory/networking? And, maybe most importantly, how to make that code be both portable between these very different architectures, and execute with high performance on all of them? 

Valentina Salapura (salapura@us.ibm.com) is an IBM Master Inventor and System Architect at the IBM T.J. Watson Research Center, Yorktown Heights, NY.

Copyright held by author.

HACC: Extreme Scaling and Performance Across Diverse Architectures

By Salman Habib, Vitali Morozov, Nicholas Frontiere, Hal Finkel, Adrian Pope, Katrin Heitmann, Kalyan Kumaran, Venkatram Vishwanath, Tom Peterka, Joe Insley, David Daniel, Patricia Fasel, and Zarija Lukić

Abstract

Supercomputing is evolving toward hybrid and accelerator-based architectures with millions of cores. The Hardware/Hybrid Accelerated Cosmology Code (HACC) framework exploits this diverse landscape at the largest scales of problem size, obtaining high scalability and sustained performance. Developed to satisfy the science requirements of cosmological surveys, HACC melds particle and grid methods using a novel algorithmic structure that flexibly maps across architectures, including CPU/GPU, multi/many-core, and Blue Gene systems. In this Research Highlight, we demonstrate the success of HACC on two very different machines, the CPU/GPU system Titan and the BG/Q systems Sequoia and Mira, attaining very high levels of scalable performance. We demonstrate strong and weak scaling on Titan, obtaining up to 99.2% parallel efficiency, evolving 1.1 trillion particles. On Sequoia, we reach 13.94 PFlops (69.2% of peak) and 90% parallel efficiency on 1,572,864 cores, with 3.6 trillion particles, the largest cosmological benchmark yet performed. HACC design concepts are applicable to several other supercomputer applications.

1. INTRODUCTION: SIMULATING THE SKY

Cosmological surveys are our windows to the grandest of all dynamical systems, the Universe itself. Scanning the sky over large areas and to great depths, modern surveys have brought us a remarkably simple, yet mysterious, model of the Universe, whose central pillars, dark matter and dark energy, point to new, and even more fundamental discoveries. The pace of progress continues unabated—the next generation of sky surveys demand tools for scientific inference that far exceed current capabilities to extract information from observations.

The already important role of cosmological simulations is expected to undergo a sea change as the analysis of surveys moves over to an approach based entirely on forward models of the underlying physics, encompassing as well the complex details of survey measurements. Such an end-to-end paradigm, based on the ability to produce realistic “universes” on demand, will stress the available supercomputing power to its limits.

The desired improvements for simulations over the next decade are often stated in terms of orders of magnitude; high accuracy and robustness are central requirements to be met by this program. Because the simulations

to be run are—and will continue to be—memory-limited on even the largest machines, stringent requirements must be simultaneously imposed on code performance and efficiency.

The rich structure of the current Universe—planets, stars, solar systems, galaxies, and yet larger collections of galaxies (clusters and filaments) all resulted from the growth of very small primordial fluctuations. These perturbations are observed today as very small temperature fluctuations in the cosmic microwave background, the fossil remnant radiation of an early hot phase of the Universe, which cooled as the Universe expanded. The primordial fluctuations grow due to the influence of gravitational attraction—this is known as the Jeans instability—although the growth is slowed by the expansion of the Universe.

A number of observations have convincingly demonstrated that only roughly a fifth of the observed matter density arises from ordinary atomic matter, the rest being a form of matter that, while it behaves normally under gravity, has exceedingly weak interactions of any other kind. This “dark matter” dominates the formation of structure—galaxies and smaller units of structure such as stars, gas clouds, and planets, all live within much larger, extended clumps of dark matter.

Given the above picture, cosmic structure formation at large scales is well described by the gravitational Vlasov-Poisson equation,¹⁷ a six-dimensional partial differential equation for the Liouville flow (1) of the phase space probability distribution function, where the gravitational potential arises self-consistently from the Poisson equation (2):

$$\partial_t f(\mathbf{x}, \mathbf{p}) + \dot{\mathbf{x}} \cdot \partial_{\mathbf{x}} f(\mathbf{x}, \mathbf{p}) - \nabla \phi \cdot \partial_{\mathbf{p}} f(\mathbf{x}, \mathbf{p}) = 0, \quad (1)$$

$$\nabla^2 \phi(\mathbf{x}) = 4\pi G a^2(t) \Omega_m \delta_m(\mathbf{x}) \rho_c. \quad (2)$$

The expansion history of the Universe is encoded in the time-dependence of the scale factor $a(t)$ governed by the cosmological model, the Hubble parameter, $H = \dot{a}/a$, G is Newton’s constant, ρ_c is the critical density (if the cosmic density is above ρ_c , the universe recollapses, if below, it expands forever), Ω_m , the average mass density as a fraction of ρ_c , $\rho_m(\mathbf{x})$ is

The original version of this paper was published in SC’13, *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis* (2013).

the local mass density, and $\delta_m(\mathbf{x})$ is the dimensionless density contrast,

$$\rho_c = \frac{3H^2}{8\pi G}, \quad \delta_m(\mathbf{x}) = \frac{(\rho_m(\mathbf{x}) - \langle \rho_m \rangle)}{\langle \rho_m \rangle}, \quad (3)$$

$$\mathbf{p} = a^2(t) \dot{\mathbf{x}}, \quad \rho_m(\mathbf{x}) = a(t)^{-3} m \int d^3\mathbf{p} f(\mathbf{x}, \mathbf{p}). \quad (4)$$

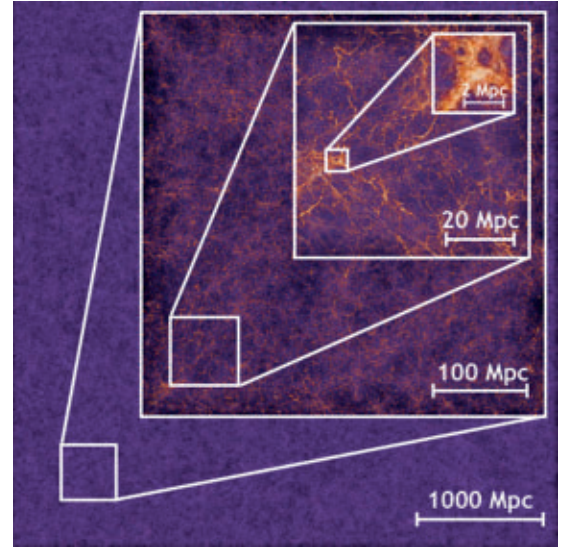
The Vlasov–Poisson equation is hopeless to solve as a PDE because of its high dimensionality and the development of nonlinear structure—including complex multistreaming—on ever finer scales, driven by the Jeans instability. Consequently, N-body methods, using tracer particles to sample $f(\mathbf{x}, \mathbf{p})$ are used; the particles follow Newton’s equations, with the forces on the particles given by the gradient of the scalar potential $\phi(\mathbf{x})$.⁴

Initial conditions are set at early times using the known properties of the primordial fluctuations. These perturbations, given by a smooth Gaussian random field, evolve into a “cosmic web” comprised of sheets, filaments, and mass concentrations called halos.^{21, 24} The first stars and galaxies form in halos and then evolve as the halo distribution also evolves by a combination of dynamics, mass accretion and loss, and by halo mergers. In addition to gravity, gasdynamic, thermal, radiative, and other processes must also be modeled. Large-volume simulations usually incorporate the latter effects via semi-analytic modeling, since the overall spatial and temporal dynamic range is too vast to be encompassed in an individual simulation.

Elementary arguments easily demonstrate the scale of the challenge to be overcome. Survey depths are of order several Gpc (1 pc = 3.26 light-years); to follow bright galaxies, halos with a minimum mass of $\sim 10^{11} M_\odot$ ($M_\odot = 1$ solar mass) must be tracked. To properly resolve these halos, the tracer particle mass should be $\sim 10^8 M_\odot$ and the force resolution should be small compared to the halo size, that is, $\sim$ kpc. This immediately implies a dynamic range (ratio of smallest resolved scale to box size) of a part in 10^6 ($\sim$ Gpc/kpc) everywhere in the *entire* simulation volume (Figure 1). In terms of the number of simulation particles required, counts range from hundreds of billions to many trillions. Time-stepping criteria follow from a joint consideration of the force and mass resolution.²⁰ Finally, stringent requirements on accuracy are imposed by the very small statistical errors in the observations—some observables must be computed at accuracies of a *fraction* of a percent.

For a cosmological simulation to be considered “high-resolution,” *all* of the above demands must be met. In addition, throughput is a significant concern. Scientific inference from cosmological observations defines a statistical inverse problem where many runs of the forward problem are needed to estimate cosmological parameters and associated errors. For such analyses, hundreds of large-scale, state of the art simulations will be required.¹² The Hardware/Hybrid Accelerated Cosmology Code (HACC) framework meets these exacting conditions in the realm of performance and scalability across a variety of node-level architectures. In this Research Highlight, we will describe

Figure 1. Zoom-in visualization of the density field in a 1.07 trillion particle, 4.25 Gpc box-size HACC simulation with 6 kpc force resolution and particle mass, $m_p \sim 5 \sim 10^8 M_\odot$. The image, taken during a late stage of the evolution, illustrates the global spatial dynamic range covered, $\sim 10^6$, although the finer details are not resolved by the visualization.



the basic ideas behind the approach, and present some representative results.

2. CURRENT STATE OF THE ART

N-body simulations in cosmology have a long history, starting with Peebles’ simulations of 300 particles in 1969,¹⁶ to today’s largest simulations evolving more than a trillion particles. Initial N^2 approaches gave way quickly to more efficient methods. Particle-mesh (PM) methods proved insufficient to obtain the required force resolution and were replaced by P^3M (Particle–Particle PM) algorithms (e.g., Ref.³), and tree codes.¹⁸ Because of the high degree of clustering in cosmological simulations, P^3M codes have been mostly displaced by tree codes (nevertheless, as demonstrated by HACC, P^3M can be resurrected for CPU/GPU systems). To localize tree walks and make handling periodic boundary conditions easier, hybrid TreePM methods were introduced, and form the mainstay of gravity-only cosmology simulations. The most popular code used today, GADGET-2,²² is also a TreePM code.

At the same time as the tree-based methods were being developed, high-resolution grid-based PM approaches using Adaptive Mesh Refinement (AMR) also made their appearance, for example, Refs.^{2, 5, 23} In the case of large-volume survey simulations, the efficiency of AMR methods is reduced because clustering occurs over the entire simulation volume, leading to high AMR computational and memory costs (in turn degrading the available force and mass resolution). Consequently, the most successful AMR applications have involved the study of a smaller number of objects such as clusters of galaxies at high resolution and with different physics modules employed. For a comparison and overview of ten different codes spanning all algorithms discussed above see Ref.¹³

Gas physics has been introduced into the structure formation codes either by using Eulerian methods following an AMR approach, or by using Smoothed Particle Hydrodynamics (SPH). Hybrid approaches are also being intensively explored. Physics at small scales (star formation, supernova feedback) is difficult to model self-consistently and is most often treated using phenomenological subgrid models. The parameters in these models are determined in part by fitting against observational data.

3. MULTI-ARCHITECTURE CHALLENGE

3.1. HACC architecture: two-level approach

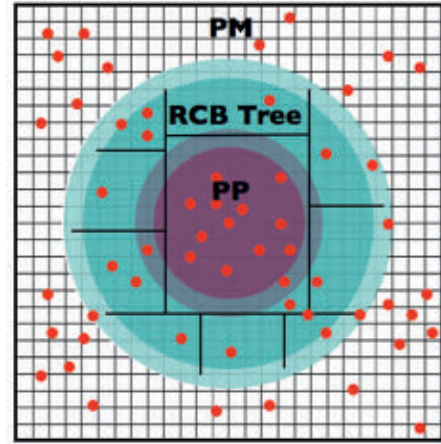
A modern code framework must confront issues raised by diverse architectures and programming models, as well as be able to respond to potentially disruptive future evolution. It should be able to gracefully incorporate multiple algorithms that interoperate with each other, to optimize them for the architecture at hand, and place minimal reliance on external resources that can potentially limit these abilities. The above remarks are a concise statement of HACC's design philosophy.

The strategy follows a two-level paradigm. As discussed in Section 2, the cosmological N-body problem is typically treated using both grid and particle-based approaches. Grid-based techniques are better suited to larger ("smooth") length scales, with particle methods having the opposite property. This suggests that the higher level of code organization should be grid-based, interacting with particle information at a lower level of the computational hierarchy. Following this central idea, HACC uses a hybrid parallel algorithmic structure, splitting the gravitational force calculation into a specially designed grid-based long/medium range spectral PM component that is essentially architecture-independent, and an architecture-adaptive particle-based short-range solver (Figure 2). The grid is responsible for four orders of magnitude of dynamic range, while the particle methods handle the critical two orders of magnitude at the shortest scales where particle clustering is maximal and the bulk of the time-stepping computation takes place.

The flexibility to respond to different nodal architectures is built into the short-range solvers; these can use direct particle-particle interactions, that is, a P^3M algorithm,¹⁵ as on Roadrunner and Titan, or use both tree and particle-particle methods as on the IBM BG/Q ("PPTreePM"). Access to multiple algorithms within the HACC framework also enables careful error testing and validation, for example, the P^3M and the PPTreePM versions agree to within 0.1% for the nonlinear power spectrum test in the code comparison suite of Ref.¹⁴

HACC's multi-algorithmic structure attacks several common weaknesses of conventional particle codes including limited vectorization, indirection, complex data structures, lack of threading, and short interaction lists. It combines MPI with a variety of nodal programming models (e.g., CUDA, OpenCL, OpenMP) to readily adapt to different platforms. HACC has been ported to conventional and Cell or GPU-accelerated clusters, Blue Gene systems, and Intel Xeon Phi systems. HACC can run at full scale on *all* available supercomputer architectures. To showcase this flexibility, we present scaling results for two

Figure 2. Informal representation of the HACC force evaluation hierarchy—(1) long/medium-range contributions from a high-order grid-based, spectrally filtered particle-mesh (PM) solver, (2) medium/short-range contributions using a (rank-local) recursive coordinate bisection (RCB) tree algorithm (green region), (3) close-range contributions using direct particle-particle (PP) interactions (magenta). Parameters governing the cross-overs are discussed in the text.



very different cases in Section 4: the IBM BG/Q systems Mira at ALCF and Sequoia at LLNL, and Titan at Oak Ridge Leadership Computing Facility (OLCF).

3.2. HACC top level

The top level of HACC's architecture consists of the domain decomposition, the medium/long-range force solver, and the interface to the short-range solver. All three aspects of this design involve new ideas to enhance flexibility and performance.

The spatial domain decomposition is in non-cubic 3-D blocks, but unlike guard zones in a typical PM method, full particle replication—"particle overloading"—is employed across domain boundaries. Overloading provides two crucial benefits. The first is that medium/long-range force calculations require no particle communication, and high-accuracy local force calculations require only sparse refreshes of the overloading zone (for details, see Refs.^{9,19}).

The second major advantage of overloading is that it frees the local force solver from handling communication tasks, which are taken care of completely at the top level. Thus new "on-node" local methods can be plugged in easily with guaranteed scalability, requiring only local optimizations. All short-range methods in HACC are local to the MPI-rank and the locality can be fine-grained further. This can be used to lower the number of levels in tree algorithms and to parallelize across fine-grained particle interaction sub-volumes. The benefits bestowed by overloading come at only modest cost: the memory overhead for a large run is only ~10%.

The long/medium range algorithm is based on a fast, spectrally filtered PM method incorporating several sophisticated features. The density field is generated from the particles using a Cloud-In-Cell (CIC) scheme¹⁵ and is then smoothed with an isotropizing spectral filter. The spectral filter reduces

the anisotropy “noise” of the CIC scheme by over an order of magnitude without recourse to inflexible higher-order spatial particle deposition methods that are more commonly used. The noise reduction allows matching of the short and longer-range forces at small scales and confers the ability to use higher-order methods, both of which have important ramifications for accuracy and performance.

The Poisson solver uses a sixth-order influence function (spectral representation of the inverse Laplacian). The gradient of the scalar potential is obtained using low-noise fourth-order Super-Lanczos spectral differencing.¹⁰ The “Poisson-solve” in HACC is the composition of all the kernels discussed above within one single discrete sum, each component of the potential gradient requiring an independent FFT. HACC uses its own scalable, high performance 3-D FFT routine implemented using a 2-D pencil decomposition (Section 4).

3.3. HACC short-range solvers

The form of the short-range force is given by subtracting the filtered grid force from the exact Newtonian force. The filtered force was determined to high accuracy using randomly sampled particle pairs and then fitted to an expression with the correct large and small distance asymptotics. Thanks to the effective use of filtering (Section 3.2), this functional form is needed only over a small, compact region, and can be represented using a fifth-order polynomial expansion, resulting in the crucial ability to vectorize computations in the main force kernel.

For heterogeneous systems such as Titan, the long/medium-range solver operates at the CPU layer. Depending on the CPU/accelerator memory balance, two different modes may be used, (1) grids held on the CPU and particles on the accelerator, or, more commonly, (2) a streaming paradigm, with grid and particle information resident in CPU memory, and short-range computations streamed through the accelerator. In both cases, the local force solve is a direct particle-particle interaction, resulting in a hardware-accelerated P³M code. The accelerated short-range algorithm on Titan outperforms the corresponding CPU-only TreePM version by more than an order of magnitude. We also have an implementation of the tree-based algorithm as used on the BG/Q, but with the tree-build and walk performed on the CPU, and the actual force evaluations performed on the GPU, leading to comparable performance as with the P³M code (currently it is a factor of two or so slower).

The streaming of particles is carried out by partitioning the 3-D particle domain into 2-D data slabs. The slab width is between 3 and 4 grid cells, as the force resolution of the top level PM solver suffices for larger scales. Therefore, the short range force calculation on one slab only requires data from adjacent slabs. We dynamically store four slabs in the GPU memory at any given time, performing the P³M algorithm on the middle slab. While the GPU performs its calculation on the chosen slab, the CPU host code simultaneously reads in the data results of the previous slab calculation, while writing the upcoming slab into GPU memory for later computation. The latency of reading and writing memory to the GPU is absorbed primarily by the GPU computation time as such memory movement can be performed simultaneously.

This memory partitioning, not only eliminates the limited memory problem of the GPU, but also drastically reduces the cost of memory movement between the CPU and the GPU, a pernicious performance chokepoint in GPU codes. In fact, each iteration (i.e., computation of the middle slab) takes longer than the simultaneous memory push, eliminating extra time spent on memory movement.

For a many-core system (e.g., BG/Q or Intel Xeon Phi), the GPU strategy is obviously not applicable, and it is more efficient to change the short-range solver to a tree-based algorithm. HACC uses a recursive coordinate bisection (RCB) tree in conjunction with a highly tuned short-range polynomial force kernel. The implementation of the RCB tree, although not the force evaluation scheme, generally follows the discussion in Ref.⁶ Two core principles underlie the high performance of the RCB tree’s design.

Spatial Locality. The RCB tree is built by recursively dividing particles into two groups, placing the dividing line at the center of mass coordinate perpendicular to the longest side of the box. Particles are then partitioned such that particles in each group occupy disjoint memory buffers. Local forces are computed one leaf node at a time. The particle data exhibits a high degree of spatial locality after the tree build; because the computation of the short-range force on the particles in any given leaf node, by construction, deals with particles only in nearby leaf nodes, the cache miss rate is extremely low.

Walk Minimization. In a traditional tree code, an interaction list is built and evaluated for each particle. The tree walk necessary to build the list is relatively slow because it involves complex conditional statements and “pointer chasing” operations. A direct N^2 force calculation scales poorly as N grows, but for a small number of particles, a thoughtfully constructed kernel can still finish the computation in a small number of cycles. The RCB tree exploits our highly tuned force kernels to reduce the overall evaluation time by shifting workload away from the slow tree-walking and into the force kernel. On many systems, tens or hundreds of particles can be in each leaf node before the critical crossover point in computational efficiency is reached.

3.4. Other features

The time-stepping in HACC is based on a 2nd order split-operator symplectic scheme that sub-cycles the short/close-range evolution within long/medium-range “kick” maps where particle positions do not change but the velocities are updated. The number of sub-cycles can vary, depending on the force and mass resolution of the simulation, from $n_c = 5$ –10. Local density estimates automatically provided by the RCB tree are used to enable adaptive time-stepping at the level of an individual leaf. HACC uses mixed precision computation—double precision is used for the spectral component of the code, whereas single precision is adequate for the short/close-range particle force evaluations and particle time-stepping.

As emphasized above, HACC’s performance and flexibility are not dependent on vendor-supplied or other high-performance libraries or linear algebra packages; the 3-D parallel FFT implementation in HACC couples high performance with a small memory footprint as compared to

available libraries. Unlike some other N-body codes that have been specially tuned for performance, no special hardware use is associated with HACC, and assembly level programming is not required.

To summarize, the HACC framework integrates multiple algorithms and optimizes them across architectures; it has several performance-enhancing features, for example, overloading, spectral filtering and differentiation, mixed precision, compact local trees, and locally adaptive time-stepping. Finally, weak scaling is a function only of the spectral solver; HACC's 2-D domain decomposed FFT guarantees excellent performance and scaling properties (Section 4).

4. PERFORMANCE

4.1. Target architectures and environments

The defining characteristic of HACC, as already discussed, is its ability to run on diverse architectures (multi/many-core as well as heterogenous) without sacrificing performance or scalability. We will showcase this on two very different architectures: the GPU-accelerated system Titan and the BG/Q systems Sequoia and Mira. These machines currently occupy rank two, three, and five in the Top 500 list (<http://www.top500.org/>). These architectures represent two very different approaches to parallel supercomputing, a smaller number of "hot" nodes with a larger flops/bandwidth imbalance (Titan) versus a larger number of lower compute intensity nodes, with a more balanced network configuration (Mira and Sequoia). It is important to note that while code ports to Titan have involved a fair degree of effort (measured in man-years), the initial transition of HACC to Titan took less than a two-person month.

Titan, a hybrid Cray XK7 system, is the third generation of major capability computing systems at the OLCF. The initial configuration was accepted in February 2012 and consisted of 18,688 compute nodes for a total of 299,008 AMD Opteron 6274 "Interlagos" processor cores and 960 NVIDIA X2090 "Fermi" Graphical Processing Units (GPU). The peak performance of the Opteron cores is 2.63 PFlops and the peak performance of the GPUs is 638 TFlops in double precision. In late 2012, the 960 NVIDIA X2090 processors were removed and replaced with 18,688 of NVIDIA's next generation Tesla K20X "Kepler" processors, with a total system peak performance in excess of 27 PFlops in double precision.

The BG/Q is the third generation of the IBM Blue Gene line of supercomputers. The BG/Q Compute chip (BQC) combines CPUs, caches, network, and a messaging unit on a single chip; each BG/Q node contains the BQC and 16GB of DDR3 memory. Each BQC uses 17 augmented 64-bit PowerPC A2 cores with specific enhancements: (1) 4 hardware threads and a SIMD quad floating point unit (Quad Processor eXtension, QPX), (2) a sophisticated L1 prefetching unit with both stream and list prefetching, (3) a wake-up unit to reduce certain thread-to-thread interactions, and (4) transactional memory and speculative execution. Of the 17 BQC cores, 16 are for user applications and one for system services. Each core has access to a private 16KB L1 data cache and a shared 32MB multi-versioned L2 cache connected by a crossbar. The A2 core runs at 1.6GHz and the QPX allows for 4 FMAs per cycle, translating to a peak performance of 204.8 GFlops

for the BQC chip. The BG/Q network has a 5-D torus topology; each node has 10 communication links with a peak total bandwidth of 40GB/s. Our results have been obtained on Sequoia, a 96 rack system (1,572,864 cores) with ~20 PFlops peak performance and on Mira, a 48 rack system (786,432 cores) with ~10 PFlops peak performance.

4.2. Performance results

The results are presented in three parts: performance and scaling of (i) the FFT and hence of the medium/long range solver, and of the full code on (ii) Titan, and (iii) on BG/Q systems. Weak and strong scaling results are shown for all cases. For the full code runs, the particle mass is $\sim 5 \cdot 10^{10} M_{\odot}$ and the force resolution, 6 kpc. All simulations are for a Λ CDM (Cold Dark Matter) model with $\Omega_m = 0.265$. Simulations of cosmological surveys focus on large problem sizes, therefore the weak scaling properties are of primary interest. The full code exhibits essentially perfect weak scaling out to 16,384 nodes of Titan (~90% of the system) at 92.2% parallel efficiency. It strong scales up to almost half of Titan on a problem with (only) 1024^3 particles.

On the BG/Q systems, HACC weak scales to the full machine size, achieving a performance of 13.94 PFlops on 96 racks, at around 67–69% of peak in most cases (up to 69.37%) at an efficiency of 90%. We demonstrate strong scaling up to one rack on a 1024^3 particle problem. Finally, the biggest test run evolved more than 3.6 trillion particles ($15,360^3$), exceeding by more than an order of magnitude the largest high-resolution cosmology run performed to date. Extensive details about the performance results on the BG/Q systems and how the high peak performance was achieved are given in Ref.⁷ Here we give a summary of those results.

FFT scaling and the Poisson solver. The weak scaling of HACC is controlled by the FFT that underlies the spectral Poisson solver (Section 3). To achieve extreme scalability, HACC has its own fast, portable, and memory-efficient pencil-decomposed, non-power-of-two FFT (data partitioned across a 2-D subgrid), allowing $N_{\text{rank}} < N_{\text{FFT}}^2$, sufficient for use in any supercomputer in the foreseeable future. The FFT is composed of interleaved transposition and sequential 1-D FFT steps, where each transposition only involves a subset of all tasks; the transposition and 1-D FFT steps are overlapped and pipelined, with a reduction in communication hotspots in the interconnect. Details of the implementation are rather complex, requiring careful scheduling of communication phases in order to avoid deadlock.

Detailed timing information for the FFT on the BG/Q and Titan is given in the original SC'13 paper. Both strong and weak scaling tests were performed. For the strong scaling test, as ranks increase from 256 to 8192 (8 ranks per node on the BG/Q and one rank per node on Titan), the scaling remains close to ideal, similar on both machines. In the second set of scaling tests, the grid size per rank is held constant, at approximately 200^3 for the BG/Q and 300^3 for Titan (the last Titan run was increased to 400^3 particles per rank). FFT scaling is demonstrated on the BG/Q up to 16 racks and to a size of $10,240^3$. The performance is remarkably stable, predicting excellent FFT performance on the largest systems and, as shown in the next section, holding up to 96 racks.

For our final full runs we measured the overall timing only—the excellent scaling of the full code is proof of the FFT scaling up to $15,360^3$ grid sizes on more than 1.5 million cores.

HACC scaling up to 16,384 nodes on Titan. We present performance data for two cases: (1) weak scaling on Titan with up to 16,384 nodes; and (2) strong scaling on up to 8,192 nodes with a fixed-size simulation problem. Timing results are obtained by averaging over 15 substeps.

Weak Scaling: We ran with 32 million particles per node in a fixed (nodal) physical volume of $(360 \text{ Mpc})^3$, representative of the particle loading in actual large-scale simulations (the GPU version of the code was run with one MPI rank per node). The results are shown in Figure 3, including timing results for a 1.1 trillion particle run, where we have kept the volume per node the same but increased the number of particles per node by a factor of two to 64.5 million. This benchmark demonstrates essentially perfect weak scaling with respect to time to solution.

Strong Scaling: Many-core based architectures are tending inexorably towards a large number of heterogeneous cores per node with a concomitant decrease in the byte/flop ratio, a defining characteristic of exascale systems. For these future-looking reasons—anticipating the strong-scaling barrier for large-scale codes—and for optimizing wall-clock at fixed problem size, it is important to establish the robustness of the strong scaling properties of the HACC algorithms.

We ran a fixed-size 1024^3 particle problem while increasing the number of nodes from 32 to 8192, almost half of Titan. The results are shown in Figure 3. For the run with the smallest number of nodes, we utilize ~30% of the available CPU memory (32 GB per node), for the largest, we use less than 1%. Up to 512 nodes, HACC strong-scales almost perfectly, after which the scaling degrades somewhat. This is not surprising—at this low particle loading the GPUs lack the computational work to hide the particle transfer penalty to the CPU. A utilization of less than 1% of the available memory is not a real-world scenario, yet even at this value, HACC performs extremely well.

HACC scaling up to 96 racks of the BG/Q. We present data for two cases: (1) weak scaling at 90% parallel efficiency with up to 1,572,864 cores (96 racks); (2) strong scaling with up to 16,384 cores with a fixed-size problem to explore future systems with lower memory per core. Timing results are obtained by averaging over 50 substeps.

Weak Scaling: We ran with 2 million particles per core, a typical particle loading in actual large-scale simulations on BG/Q systems. Tests with 4 million particles per core produce very similar results. As demonstrated in Figure 4, weak scaling is ideal up to 1,572,864 cores (96 racks), where HACC attains a peak performance of 13.94 PFlops and a time per particle per substep of ~0.06 ns for the full high-resolution code. This problem, with 3.6 trillion particles, is the largest cosmological benchmark ever performed. The time to solution is set by the science use requirement, that is, running massive high-precision HACC simulations on a production basis—within days rather than weeks. The performance achieved allows runs of 100 billion to trillions of particles in a day to a week of wall-clock time.

Figure 3. Weak and strong scaling on Titan. Weak scaling is reported for ~32 million particles per node. The time per substep per particle is shown as a function of the number of nodes: The performance and time to solution demonstrate essentially perfect scaling (black line). Strong scaling results are for a fixed-size problem— 1024^3 particles in a 1.42 Gpc box. The final number of nodes is 8192, approximately half of Titan. Recent improvements in absolute performance are also shown (see Section 5.1).

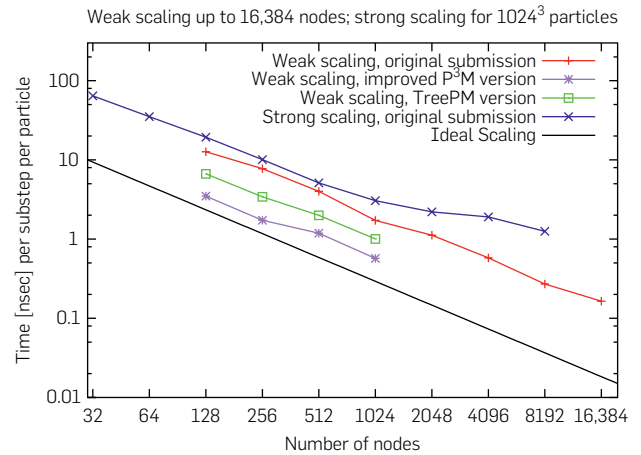
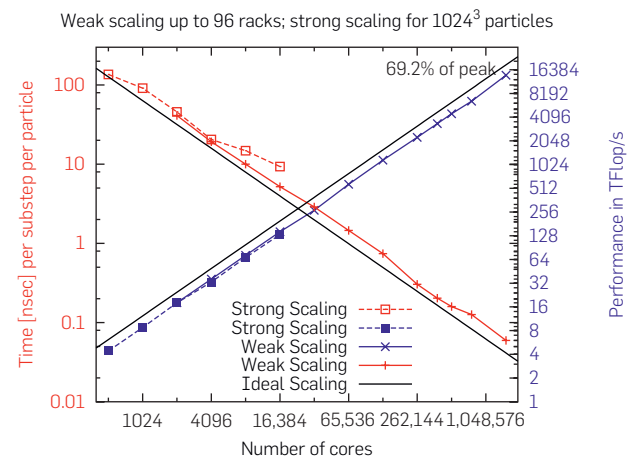


Figure 4. Weak and strong scaling on the BG/Q; time per substep per particle (red) and overall performance (blue), as a function of the number of cores. The offset black lines indicate ideal scaling. Weak scaling (solid lines with crosses) is reported for ~2 million particles per core for up to 96 racks. Performance and time to solution demonstrate essentially perfect scaling. Strong scaling (dashed lines, boxes) follows the same set up as for Titan (Figure 3), the number of cores going from 512 to 16,384. The timing scales nearly perfectly to 8192 cores, then degrades slightly; performance stays high throughout.



Strong Scaling: The problem set-up follows that on Titan, using 512 to 16,384 cores, and spanning a per node memory utilization of ~57%, a typical production run value, to as low as 7%. The actual utilization scales by a factor of 8, instead of 32, because on 16,384 nodes we are running a very small simulation volume per rank with high overloading memory and compute cost. Given that the algorithms are designed to run at >50% of per node memory utilization to a factor of 4 less (~15%), the strong scaling, as depicted in Figure 4

is impressive. It stays near-ideal throughout, slowing down at 16,384 cores, only because of extra computations in the overloaded regions. This test demonstrates that the HACC algorithms will work well in situations where the byte/flop ratio is significantly smaller than the optimal plateau for the BG/Q.

5. UPDATES AND FUTURE EVOLUTION

5.1. HACC updates

After the original SC'13 paper (on which this Research Highlight is based) appeared, a number of improvements have been implemented that further increase the efficiency of the short-range solvers. These include improved data streaming to the GPU, a new task-based load-balancing scheme, multiple RCB trees/rank, and locally adaptive time-stepping.⁸

On Titan, we determined that PCI bus latencies were best avoided by pushing data in larger blocks, as opposed to asynchronously pushing 2-D slabs. The new code calculates the maximum amount of memory that can be allocated on the GPU and pushes data of that size. With a larger data size, the calculation time is increased on the device, reducing the overall time spent moving data between the host CPU and GPU. The kernel itself was also updated and rewritten to lower the register pressure per thread, allowing for an occupancy of 100%, maximally hiding latencies in memory fetching. Other improvements include further loop unrolling, and actual alterations in the assembly code. The particle interaction algorithm itself was unaltered, but by analyzing various aspects of the PTX assembly, the compiled kernel was further optimized. This led to a three times to four times improvement.

The principle behind the load balancing technique is to partition each node volume into a set of overlapping data blocks, which contain “active” and “passive” particles—analogue to the overloading particle scheme between nodes. Each block can independently perform the short-range force calculations on its data, where it correctly updates the interior active particles, and streams the bounding passive particles. In this form, one can picture each node as a separate HACC simulation, and the data blocks are the equivalent nodal volume decompositions with overloading zones. The scheme to perform a short-range force timestep is as follows: (1) Each node partitions itself into overlapping data blocks, (2) evolves the blocks independently, and (3) reconciles the active particles, whereby removing the unnecessary duplicated passive ones. Once the simulation data has been subdivided into smaller independent work items, these blocks can be communicated to any nodes that have the available resources to handle the extra workload. Load-balancing can now be performed; more details are given in Ref.¹¹

On the BG/Q, to increase the amount of parallel work in the short-range solver, HACC builds multiple RCB trees per rank. First, the particles are sorted into fixed bins, where each bin is roughly the length-scale of the short-range force. An RCB tree is constructed within each bin, and because this process is independent of all other bins, it can be done in parallel, providing a significant performance boost to the overall force computation. Note that when the force on the particles in each leaf node is computed, not only must the

parent tree be searched, but so must the other 26 neighboring trees. Only nearest neighbors need to be considered because of the limited range of the short-range force. While searching neighboring trees adds computational expense, the trees are individually not as deep, and so the resulting walks are less expensive. Also, because we distribute “(leaf node, neighboring tree)” pairs among the threads, this scheme also increases the amount of available parallelism post-tree-build (which helps with thread-level load balancing). Overall this technique provides a significant performance advantage over using one large tree for the entire domain.⁸

5.2. HACC future evolution

HACC framework development is tightly coupled to future architectures; in particular the two-level design naturally maps the lower level (short-range solvers, and, in the very near future, particle-based hydro-solvers) to the individual node architecture. Because of this feature, architecture co-design with HACC kernels is particularly straight-forward. Conversely, knowledge of the nodal architecture allows for ease of optimization and algorithmic choices within the framework.

The choice of programming models is also connected to this two-level structure. It appears very likely that in the near future (“pre-exascale” systems), the number of nodes—distinct pieces of hardware directly connected to the main system network—will not be too different from 100,000 (or smaller). HACC has already demonstrated running at more than 1.5 million MPI ranks, and it is very unlikely that a significant improvement in this number will be needed. Programming models at the node level will evolve, and, as already demonstrated, HACC can easily adapt to these.

Future nodes will express unprecedented levels of concurrency; possibly thousands of independent threads. The grand challenge for pre-exascale applications will be how to adapt to this change. In principle, HACC has the potential to utilize a large number of independent streams effectively; precisely how to do this is a key focus area. In terms of the bytes/flop ratio, given a pre-exascale system with notional numbers of 100 PFlops and 10 PB of total memory, the results presented here show that HACC is ready for such a machine today.

There are a number of technical issues such as complex memory hierarchies (e.g., NVRAM) and bandwidth, power management, and resilience technologies that we continue to monitor. HACC is one of the benchmark codes for gathering information regarding future supercomputers at Argonne, Livermore, and Oak Ridge. This provides an opportunity to stay abreast of the latest advances; because HACC does not rely on “black box” libraries or packages, it retains the key advantage of allowing optimization to be a continuous process.

Many of the ideas and methods presented here are relatively general and can be re-purposed to benefit other HPC applications, especially in areas such as accelerator beam dynamics and plasma simulations, particle transport codes, and for molecular dynamics simulations.

An important aspect of HACC not mentioned so far is an associated parallel in situ analysis and visualization framework called CosmoTools that runs in tandem with HACC simulations to perform “on the fly” analysis (computation of

summary statistics, halo finding, halo merger trees, etc.) and data reduction tasks. The very large simulations undertaken by the HACC framework make having such a capability an absolute requirement, as extensive post-processing of the raw data outputs is almost impossible to carry out. Future development of HACC will also involve associated development of CosmoTools.

Early science results from a number of HACC simulations include a suite of 64 billion particle runs for baryon acoustic oscillations predictions for Baryon Oscillation Spectroscopic Survey^a (BOSS) carried out on Roadrunner²⁵ and a high-statistics study of galaxy cluster halo profiles.¹ This has been followed by some very large simulations, among them the largest high-resolution N-body runs in cosmology to date. These include a 1.1 trillion particle simulation (Figure 1)⁸ run on Mira, a simulation with roughly half this number of particles, but an order of magnitude better mass resolution run on Titan,¹¹ and a suite of thirty 64 billion particle simulations for the BOSS survey spanning multiple cosmologies, and designed to construct full-sky synthetic catalogs out to a redshift depth of $z \sim 0.8$. The large Mira run has been used to generate synthetic galaxy catalogs for the next-generation Dark Energy Spectroscopic Instrument (DESI)^b and results from both of the large runs will be used to construct synthetic skies for the Large Synoptic Survey Telescope (LSST).^c

Acknowledgments

We are indebted to Bob Walkup for running HACC on a prototype BG/Q system at IBM and to Dewey Dasher for help in arranging access. At ANL, we thank Susan Coghlan, Paul Messina, Mike Papka, Rick Stevens, and Tim Williams for obtaining allocations on different Blue Gene systems. At LLNL, we are grateful to Brian Carnes, Kim Cupps, David Fox, and Michel McCoy for providing access to Sequoia. At ORNL, we thank Bronson Messer and Jack Wells for assistance with Titan. This research used resources of the ALCF, which is supported by DOE/SC under contract DE-AC02-06CH11357 and resources of the OLCF, which is supported by DOE/SC under contract DE-AC05-00OR22725.

^a <http://www.sdss.org/surveys/boos/>.

^b <http://desi.lbl.gov/>.

^c <http://www.lsst.org/lsst/>.

References

- Bhattacharya, S., Habib, S., Heitmann, K., Vikhlinin, A. Dark matter Halo profiles of massive clusters: Theory versus observations. *Astrophys. J.* 766 (2013), 32.
- Bryan, G.L., Norman, M.L. In *12th Kingston Meeting on Theoretical Astrophysics, Proceedings of Meeting Held in Halifax, Nova Scotia (ASP Conference Series # 123)*, D.A. Clarke and M. Fall, eds. 1996; see also O'Shea, B.W., Nagamine, K., Springel, V., Hernquist, L., Norman, M.L., *Astrophys. J. Supp.* 160 (2005), 1.
- Couchman, H.M.P., Thomas, P.A., Pearce, F.R. Hydra: An adaptive-mesh implementation of P 3M-SPH *Astrophys. J.* 452, 797 (1995).
- For a review of cosmological simulation methods, see also Dolag, K., Borgani, S., Schindler, S., Diaferio, A., Bykov, A.M. *Space Sci. Rev.* 134 (2008), 229.
- Fryxell, B., et al. FLASH: An adaptive mesh hydrodynamics code for modeling astrophysical thermonuclear flashes. *Astrophys. J. Supp.* 131 (2000), 273.
- Gafton, E., Rosswog, S. A fast recursive coordinate bisection tree for neighbour search and gravity. *Mon. Not. R. Astron. Soc.* 418 (2011), 770.
- Habib, S., Morozov, V., Finkel, H., Pope, A., Heitmann, K., Kumaran, K., Peterka, T., Insley, J., Daniel, D., Fasel, P., Frontiere, N., Lukić, Z. arXiv:1211.4864, *Supercomputing 2012*.
- Habib, S., Pope, A., Finkel, H., Frontiere, N., Heitmann, K., Daniel, D., Fasel, P., Morozov, V., Zagaris, G., Peterka, T., Vishwanath, V., Lukić, Z., Sehrish, S., Liao, W.-K. HACC: Simulating sky surveys on state-of-the-art supercomputing architectures. *New Astron.* 42 (2016), 49 arXiv:1410.2805 [astro-ph.IM].
- Habib, S., Pope, A., Lukić, Z., Daniel, D., Fasel, P., Desai, N., Heitmann, K., Hsu, C.-H., Ankeny, L., Mark, G., Bhattacharya, S., Ahrens, J. Hybrid petacomputing meets cosmology: The Roadrunner Universe project. *J. Phys. Conf. Ser.* 180 (2009), 012019.
- Hamming, R.W. *Digital Filters*. Dover, Publications, Mineola, New York 1998.
- Heitmann, K., Frontiere, N., Sewell, C., Habib, S., Pope, A., Finkel, H., Rizzi, S., Insley, J., Bhattacharya, S. The Q continuum simulation: Harnessing the power of GPU accelerated supercomputers. *J. - Astrophys. J. Supp.* 219 (2015), 34 arXiv:1411.3396 [astro-ph.CO].
- Heitmann, K., Higdon, D., White, M., Habib, S., Williams, B.J., Lawrence, E., Wagner, C. The Coyote universe. II. Cosmological models and precision emulation of the nonlinear matter power spectrum *Astrophys. J.* 705 (2009), 156.
- Heitmann, K., Lukić, Z., Fasel, P., Habib, S., Warren, M.S., White, M., Ahrens, J., Ankeny, L., Armstrong, R., O'Shea, B., Ricker, P.M., Springel, V., Stadel, J., Trac, H. The cosmic code comparison project. *Comput. Sci. Dis.* 1 (2008), 015003.
- Heitmann, K., Ricker, P.M., Warren, M.S., Habib, S. Robustness of cosmological simulations. I. Large-scale Structure. *Astrophys. J. Supp.* 160 (2005), 28.
- Hockney, R.W., Eastwood, J.W. *Computer Simulation Using Particles*. Adam Hilger, New York, 1988.
- Peebles, P.J.E., Structure of the coma cluster of galaxies. *Astron. J.* 75 (1970), 13.
- Peebles, P.J.E. *The Large-Scale Structure of the Universe*. Princeton University Press, Princeton, New Jersey 1980.
- Pfalzner, S., Gibbon, P. *Many-Body Tree Methods in Physics*. Cambridge University Press, 1996; see also Barnes, J., Hut, P. *Nature* 324, 446 (1986); Warren, M.S., Salmon, J.K. Technical Paper, Supercomputing, Cambridge University Press, New York, USA 1993.
- Pope, A., Habib, S., Lukic, Z., Daniel, D., Fasel, P., Desai, N., Heitmann, K. *Comput. Sci. Eng.* 12 (2010), 17.
- The accelerated universe. Power, C., Navarro, J.F., Jenkins, A., Frenk, C.S., White, S.D.M., Springel, V., Stadel, J., Quinn, T. The inner structure of Λ CDM haloes – I. A numerical convergence study. *Mon. Not. R. Astron. Soc.* 338 (2003), 14.
- Shandarin, S.F., Zeldovich, Ya.B. The large-scale structure of the universe: Turbulence, intermittency, structures in a self-gravitating medium. *Rev. Mod. Phys.* 61 (1989), 185.
- Springel, V. The cosmological simulation code GADGET-2. *Mon. Not. R. Astron. Soc.* 364 (2005), 1105.
- Teyssier, R. Cosmological hydrodynamics with adaptive mesh refinement. A new high resolution code called RAMSES. *A&A* 385 (2002), 337.
- White, M. The mass of a halo. *Astron. and Astrophys.* 367 (2001), 27.
- White, M., Pope, A., Carlson, J., Heitmann, K., Habib, S., Fasel, P., Daniel, D., Lukić, Z. Particle mesh simulations of the Ly α forest and the signature of Baryon acoustic oscillations in the intergalactic medium. *Astrophys. J.* 713 (2010), 383.

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

Salman Habib, Vitali Morozov, Nicholas Frontiere, Hal Finkel, Adrian Pope, Katrin Heitmann, Kalyan Kumaran, Venkatram Vishwanath, Tom Peterka, and Joe Insley ([habib, morozov, nfrontiere, hfinkel, apope, heitmann, kumaran, venkat, tpeterka, insley]@anl.gov), Argonne National Laboratory, Lemont, IL.

David Daniel and Patricia Fasel ([ddd, pfj]@lanl.gov), Los Alamos National Laboratory, Los Alamos, New Mexico.

Zarija Lukić (zarija@lbl.gov), Lawrence Berkeley National Laboratory, Berkeley, CA.

Bowling Green State University **Shantanu and Reni Narayan Professorship** **in Computer Science (Associate or Full** **Professor)**

The Department of Computer Science is seeking an associate or full professor with expertise in the area of Software Engineering. Successful candidates shall have a record of exceptional teaching, experience in developing extracurricular experiences (internships, student groups), and ability to lead and mentor new faculty in educational activities. The Narayan Professorship was established to attract and retain distinguished educators and to encourage teaching excellence in the classroom.

Area of specialization is software engineering, including but not limited to: software testing and quality assurance, software architecture and design, usability engineering, and software verification. Applicants must hold a Ph.D. in CS (or closely related field), and be committed to excellence in teaching, scholarly research, and external funding.

BGSU offers a small town atmosphere with easy access to Columbus, Detroit, and Ann Arbor. BGSU is an AA/EEO/Vet employer. We encourage applications from women, minorities, veterans, and individuals with disabilities regardless of age, gender identity, genetic information, religion, or sexual orientation. Email a letter of interest, along with curriculum vitae, statement of teaching philosophy and research agenda, contact information for three professional references, and selected examples of teaching and scholarly work by Sunday, January 8, 2017 to cs-search@bgsu.edu.

We will contact your references. We will select a small number of finalists to come to campus for an interview. An official transcript of the terminal degree and a background check are also required for employment. For details, go to <http://www.bgsu.edu/arts-and-sciences/computer-science.html/jobs>

California State University, Fullerton

The Department of Computer Science invites applications for tenure-track positions at the **Assistant Professor** level starting August 2017. For a complete description of the department, the position, desired specialization and other qualifications, please visit http://hr.fullerton.edu/diversity/job-openings/ft/9172BR_computer_science.asp

Carnegie Mellon University **School of Computer Science**

The School of Computer Science at Carnegie Mellon University Carnegie Mellon University is seeking faculty candidates for teaching track positions. Individuals seeking this position will be responsible for leading the growth of our curricu-

lum, including improving educational outcomes and extending our reach.

Teaching track faculty are responsible for teaching courses as well as overseeing broader aspects of the educational program, including curriculum development, student advising, outreach, undergraduate research, and departmental service. You should have a Ph.D., a background of demonstrated excellence and dedication to teaching, and must be prepared to teach undergraduate lecture courses. You will also work with the existing faculty to improve and revise the curriculum. We seek candidates with demonstrated excellence in teaching and commitment toward building an equitable and diverse scholarly community, and especially invite candidates with a demonstrated track record in mentoring and engaging students from groups traditionally underrepresented in computer science.

The School of Computer Science consists of seven departments, spanning a wide range of topics in computer science and the application of computers to real-world systems. The positions are specific to each department, though in certain cases, joint positions are also possible.

Application Instructions:

<https://webapps.cs.cmu.edu/FacultyApplication/SCS-Teaching/Welcome>

Deadline for Applications is January 3, 2017.

In your cover letter, please indicate clearly the department you are applying to.

For more information of the specific departments, please see below:

Machine Learning Department:

As the world's only academic Machine Learning Department, we occupy a unique position in defining the standard curriculum for the field of machine learning – one that is used as a template by many other universities. With the increasing importance of machine learning over recent years, our course enrollments have more than doubled, and requests for us to service students beyond our local campus have grown significantly. For more information, please see: http://www.ml.cmu.edu/Faculty_Hiring.html

The Language Technologies Institute:

The Language Technologies Institute is an academic department dedicated to the study of human language and information technologies. Its faculty perform groundbreaking research in the areas of Natural Language Processing, Computational Linguistics, Information Extraction, Summarization & Question Answering, Information Retrieval, Text Mining & Analytics, Knowledge Representation, Reasoning & Acquisition, Language Technologies for Education, Machine Learning, Machine Translation, Multimodal Computing and Interaction, Speech Processing, and Spoken Interfaces & Dialogue Processing. For more information, please see: <http://lti.cs.cmu.edu/teaching-track-faculty-position>

The Institute for Software Research:

The Institute for Software Research is an academic department dedicated to the study of software engineering and societal computing, with approximately thirty faculty members. Professional experience in software engineering and familiarity with current software engineering technologies and methods is desirable. For more information, please see: <http://www.isri.cmu.edu/jobs/teaching-track.html>

Computer Science Department:

Teaching track faculty in the computer science department are able to work with some of the brightest undergraduate students in the nation and with leading faculty in the recently opened Gates Hillman Center. The department is committed to diversity in education, with the 2016 incoming undergraduate class including 48% women. For more information, please see: <https://csd.cs.cmu.edu/content/faculty-hiring>

Carnegie Mellon University **Human-Computer Interaction Institute**

The Human-Computer Interaction Institute at Carnegie Mellon University is hiring tenure-track faculty this year. We are looking for a wide range of expertise. More information can be found at: <http://hcii.cmu.edu/careers/2016/carnegie-mellon-hcii-hiring-tenure-track-faculty>.

Concordia University

Concordia University's Faculty of Engineering and Computer Science hosts over 7500 undergraduate/graduate students, and prepares the next generation of technical leaders and entrepreneurs to address complex real-world problems. We offer a multi-disciplinary and research-engaged environment dedicated to incubating innovation, excellence and success. Our teaching and research is daring and transformative and contributes significantly to a sustainable intellectual and economic development of our community. We connect ideas with people and we are redefining the university experience. The Faculty seeks outstanding candidates for:

Department of Computer Science and Software Engineering

- A Natural Sciences and Engineering Research Council (NSERC) Canada Research Chair Tier I (Full Professor) in the field of Software Engineering.
- One tenure-track position at the rank of Assistant Professor in the area of programming languages.
- One tenure-track position at the rank of Assistant Professor in the area of big data analytics.

Concordia Institute for Information Systems Engineering

- One tenure-track position at the rank of Assistant Professor in the area of Smart Grid Security, Control

Systems Security, Cyber-Physical Systems, Security and Critical Infrastructure Protection.

For detailed information about working at Concordia, these positions, and deadlines, visit <http://www.concordia.ca/encs/about/jobs.html>.

All qualified candidates are encouraged to apply; however, Canadians and Permanent Residents will be given priority. Concordia is strongly committed to employment equity within its community, and to recruiting a diverse faculty and staff. The University encourages applications from all qualified candidates, including women, members of visible minorities, Aboriginal persons, members of sexual minorities, persons with disabilities, and others who may contribute to diversification.

Duke University

The Department of Computer Science at Duke University in Durham, North Carolina, invites applications and nominations for a Professor of the Practice position starting in July 2017. The appointment will be made at the Assistant/Associate/Full Professor of the Practice rank depending on the candidate's experience. Duke has a long history of supporting faculty with practice-of titles who are educators, practitioners, and scholars, and the Computer Science department has set high standards in providing a teaching and work environment in which practice-of faculty thrive.

The ideal candidate will have a strong commitment to and demonstrated excellence in teaching. Successful candidates at the Associate and Full Professor of the Practice level will

show examples of academic, scholarly, and educational success outside the classroom that have resulted in, or have the potential to result in, significant advancements in computer science and engineering education. Successful candidates must show commitment to educating a broad and diverse group of students and in working to increase the participation and success of students from groups underrepresented in computer science.

The term of an initial appointment depends on the rank, but is typically four years for a junior position. Reappointment and promotion are governed by departmental and university bylaws and guidelines that encourage faculty to be innovative, to succeed locally and nationally, and to develop new courses and curricula that mesh with ongoing and new initiatives. Practice-of faculty typically teach two courses per semester with graduate and undergraduate TA support. We especially encourage faculty with interest and experience in mobile computing to apply.

Candidates should submit, via online application to Academic Jobs Online:

1. a short cover letter describing their background and interests in teaching undergraduates at Duke University and in promoting inclusion and diversity in computer science,
2. a curriculum vita,
3. a statement regarding teaching philosophy and experience, including evidence of commitment to promoting inclusion and diversity as well as teaching evaluations or other evidence of teaching effectiveness,
4. a statement of scholarly activities (which may include research on pedagogical focus), and

5. contact information for at least three referees who can comment on the candidate's pedagogical skills, among other attributes (provide referee name, title, address and email).

Fordham University

Department of Computer and Information

Science- Multiple Openings

Department Chair (Full Professor)

Two Tenure Track Assistant Professor Positions Lecturer & Postdoc Positions

The Department of Computer & Information Science at Fordham University is undergoing a rapid expansion in both its undergraduate and graduate programs, and Fordham's administration is committed to increasing the department's local and national distinction in research and science education. The department offers undergraduate programs at Fordham's Rose Hill campus in the Bronx and Lincoln Center campus in Manhattan. The department also offers graduate degrees in Computer Science, Data Analytics, and Cybersecurity at its Manhattan campus.

The department invites applications for department chair, assistant professors, and lecturer/postdocs, to start fall 2017. The chair position requires distinguished research, an excellent record of external funding, significant administrative experience, and strong experience in program development, collaborative academic leadership, and teaching. The tenure track positions require a Ph.D. in Computer Science or a related field, a commitment to teaching excellence, and an active program of research with the potential to attract external research funding. One of the assistant professor positions must be in the area of cybersecurity, but applications in all areas of CIS are encouraged for the second position, although applicants in the area of database systems are especially encouraged to apply. The department also has several positions for lecturers and postdocs. Lecturer positions (MS required, Ph.D. preferred) are one year renewable with a 4-4 teaching load and postdoc positions are renewable for up to three years with a 3-3 teaching load.

Applications can be electronically submitted to Interfolio Scholar Services:

For Department Chair position: apply. interfolio.com/39175.

For Cybersecurity Assistant Professor position: apply. interfolio.com/38909

For second Assistant Professor position: apply. interfolio.com/38910

For Lecturer and Postdoc positions: apply. interfolio.com/39204.

Include (1) Cover letter with qualifications, (2) Curriculum vitae, (3) Teaching Statement, (4) at least three letters of recommendation, (5) Research Statement, and (6) Sample scholarship (lecturer applicants may omit the last two items). Chair applicants should additionally include a brief statement of administrative philosophy. Applications will be accepted until the position is filled but preference will be given to applications received by January 1, 2017.

For more information about the department and the open positions, visit <http://www.cis.fordham.edu> or contact Palma Hutter at hutter@fordham.edu. Chair applicants may contact Dr. Gary Weiss at gaweiss@fordham.edu for



香港中文大學 The Chinese University of Hong Kong

Dean of the Faculty of Engineering

Founded in 1963, The Chinese University of Hong Kong (<http://www.cuhk.edu.hk>) is a forward-looking and intellectually vigorous university with the mission to be a first-class comprehensive research university, regionally and internationally. With a team of over 3,000 full-time teaching and research staff, the University offers a broad spectrum of programmes up to the PhD level in various disciplines organized under eight Faculties (namely Arts, Business Administration, Education, Engineering, Law, Medicine, Science and Social Science). In 2015-16, the undergraduate and postgraduate enrolments in the University's publicly-funded programmes have reached 16,500 and 3,500 respectively.

The Faculty of Engineering (<http://www.erg.cuhk.edu.hk/>) comprises the Departments of Computer Science and Engineering, Electronic Engineering, Information Engineering, Mechanical and Automation Engineering, and Systems Engineering and Engineering Management. The Faculty has about 300 full-time teaching and research staff, 2,300 undergraduate and 640 postgraduate research students.

The University now invites applications and nominations of qualified candidates for the Deanship of the Faculty. The Dean will be a member of the University senior management team, reporting to the University Council via the Vice-Chancellor/President or the Provost. As the academic and executive head of the Faculty, the Dean will provide academic leadership and discharge administrative responsibilities in respect of academic, staff, resource (budget and space) as well as student matters. He/she will also actively engage in alumni and community relations and in extending networks.

Candidates should have an excellent academic standing appropriate for appointment at the level of a full Professor in the Faculty. They should have an appreciation of the breadth of research/educational developments in the relevant fields and the range of intellectual interests represented in the Faculty, demonstrated capability of academic leadership and strategic management in higher education institutions, a long-term vision for the development of the Faculty, and excellent interpersonal and communication skills.

Salary and fringe benefits for the post will be highly competitive, commensurate with qualifications and experience.

Please send applications/nominations under confidential cover to the Search Committee for the Dean of the Faculty of Engineering, c/o Office of the Vice-Chancellor/President, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong [fax: (852) 3942 0947; e-mail: SCDeanship-Engg@cuhk.edu.hk]. All applications/nominations will be treated in strict confidence. The University's Personal Information Collection Statement will be provided upon request.

Consideration of applications/nominations will continue until the post is filled. The University reserves the right to fill the post by invitation.

additional information.

Fordham is an independent, Catholic University in the Jesuit tradition that welcomes applications from all backgrounds. Fordham is an equal opportunity employer.

George Mason University

The George Mason University Department of Computer Science, in the Volgenau School of Engineering, invites applications for tenure-track faculty positions beginning Fall 2017.

Required Qualifications

Applicants must have received a Ph.D. in Computer Science or a related field by the start date of the position, and should have demonstrated potential for excellence and productivity in research, and a commitment to high quality teaching. Exceptionally strong senior candidates may also be considered, and must have an established record of outstanding research and excellent teaching. Such candidates will be eligible for tenured Associate Professor or Professor positions.

Preferred Qualifications

While applicants in all areas of computer science will be given serious consideration, we are particularly interested in candidates in the areas of machine learning, artificial intelligence, robotics, databases, data analytics, and data-intensive computing.

About the Department

The department has over 40 faculty members with wide-ranging research interests including artificial intelligence, algorithms, autonomic computing, computational biology, computer graphics, computer vision, databases, data mining, parallel and distributed systems, real-time systems, robotics, security, software engineering, and wireless and mobile computing. The CS department has over \$6 Million in annual research funding and has 11 recipients of NSF's prestigious CAREER awards.

In addition to BS, MS and PhD programs in Computer Science, the department offers MS programs in Information Systems, Information Security and Assurance, and Software Engineering. The department also participates in an interdisciplinary MS in Data Analytics Engineering offered by the Volgenau School of Engineering. For more information on the department, visit our Web site: <http://cs.gmu.edu/??>

George Mason University is the largest public research university in Virginia, with an enrollment of over 35,000 students studying in over 200 degree programs. Mason is an innovative, entrepreneurial institution with national distinction in a range of academic fields. It was classified as an R1 research institution in 2016 by the Carnegie Classifications of Institutes of Higher Education, and was ranked number one in the 2013 U.S. News and World Report "Up-and-Coming" list of national universities. Mason is located in Fairfax in Northern Virginia at the doorstep of the Washington, D.C., metropolitan area, with unmatched geographical access to a number of federal agencies and national laboratories. Northern Virginia is also home to one of the largest concentrations of high-tech firms in the nation, providing excellent opportunities for interaction with industry.

Fairfax is consistently rated as being among the best places to live in the country, and has an outstanding local public school system.

For full consideration please submit application and application materials on-line at <http://jobs.gmu.edu> for position number F105AZ. To apply, you will need a statement of professional goals including your perspective on teaching and research (to attach as 'Other Doc'), a complete C.V. with publications, and the names of three professional references. The review of applications will begin February 1, 2017 and continue until the position is filled.

George Mason University is an equal opportunity/affirmative action employer, committed to promoting inclusion and equity in its community. All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, gender identity, sexual orientation, national origin, disability, or protected veteran status.

Indiana University School of Informatics and Computing Tenure-Track Assistant Professor

The Indiana University School of Informatics and Computing, Indianapolis, invites applications for a tenure-track assistant professor position in data science, beginning August, 2017 in the Department of Human-Centered Computing. Candidates must demonstrate an outstanding scholarly record of research, exhibited by high-impact peer-reviewed publications and a research agenda that will secure competitive, external funding.

We seek an exceptional researcher in data science. All areas of data science will be considered including data mining, statistical machine learning, descriptive, predictive, and prescriptive analytics, cloud computing, distributed databases, high performance computing, data visualization, or other areas involving the collection, organization, management, and extraction of knowledge from massive, complex, heterogeneous datasets.

The department is home to a dynamic and interdisciplinary group of faculty and students across its Data Science, Human-Computer Interaction, Informatics, and Media Arts and Science degree programs. The department has a strong emphasis on the human-centered aspects of data science, including interactive and multimodal visualizations, usable information representation and data manipulation, interactive tools for next-generation knowledge discovery, interactive environments for big data, human factors in data science, and human-computer interaction issues surrounding data science applications and decision-making processes.

The Data Science program is uniquely positioned to drive research in big data around health information, both from the patient's side and from the provider's side. Two major growing areas include predictive population analytics for health monitoring and prevention and novel paradigms for health literacy and information delivery based on user's data searches and online behavior. Faculty conduct groundbreaking, externally-funded research in a variety of areas, including emerging media technologies, human-computer interaction design, ubiquitous computing, accessibility, human-robot interaction and android science, and healthcare user interfaces.

Visit <http://indiana.peopleadmin.com/postings/3077> for full application instructions.

Please submit your application materials by January 1, 2017. However, the positions will remain open until filled.

Questions can be directed to Dr. Davide Bolchini at dbolchin@iupui.edu

The School of Informatics and Computing is eager to consider applications from women and minorities. Indiana University is an Affirmative Action/Equal Opportunity Employer. IUPUI is an Affirmative Action/Equal Opportunity Institution M/F/D/V.

Indiana University School of Informatics and Computing Department of BioHealth Informatics

The Indiana University School of Informatics and Computing at Indianapolis invites applications for an open-rank tenure-track or tenured faculty position in the Department of BioHealth Informatics (BHI). The appointment will begin August 1, 2017 at the Indiana University-Purdue University Indianapolis (IUPUI) campus. Exceptional researchers are being sought to join our fast-growing department. We welcome applications from established researchers with collaborative research teams. Candidates must demonstrate an outstanding scholarly record of research, exhibited by high-impact peer-reviewed publications and a forward-looking, vigorous research agenda that will secure competitive, external funding.

We seek candidates in all areas of Health and Biomedical Informatics, but particularly those with strong research and teaching experience in:

1. Community, consumer health and social informatics;
2. Health Information Technology and Health Information Exchange;
3. Learning healthcare system, including clinical intelligence, translational biomedical and clinical informatics; and
4. Big data, data analytics, text analytics, and data science in Health and Biomedical Informatics.

By strengthening or complementing the faculty research in the department, the ideal candidates will use creative, innovative approaches and technologies to address fundamental BHI challenges with broader societal impact, and have the potential to leverage the strengths of the IUPUI campus, including its unique location in downtown Indianapolis, interdisciplinary and collaborative environment and nation-wide leadership in the health and life sciences.

As tenure track faculty for BHI, the new faculty will have responsibility for developing their research programs and contributing to the educational and service roles of the department. The new faculty also will work closely with Department Chair, Dr. Huanmei Wu, to develop and expand the new department over the next several years.

Visit <https://indiana.peopleadmin.com/postings/2754> for full application instructions. Questions pertaining to this position may be directed to the Assistant to the Chair, Robyn Hart at rohart@iupui.edu.

The School of Informatics and Computing is eager to consider applications from women, veterans and minorities. Indiana University is an Affirmative Action/Equal Opportunity Employer. IUPUI is an Affirmative Action/Equal Opportunity Institution M/F/D/V.

Iowa State University

The Department of Computer Science in the College of Liberal Arts and Sciences at Iowa State University is seeking outstanding candidates for the endowed Lanh and Oanh Nguyen Chair in Software Engineering at the rank of Associate or Full Professor.

Responsibilities include teaching undergraduate and graduate courses; mentoring students; sustaining a strong, externally funded research program in software engineering; publishing in top tier venues; and enhancing Iowa State University's existing strengths through professional and institutional service.

Iowa State University is classified as a Carnegie Foundation Doctoral/Research University-Extensive, a member of the Association of American Universities (AAU), and ranked by U.S. News and World Report as one of the top public universities in the nation. 36,600 students are enrolled, and served by over 6,200 faculty and staff (see www.iastate.edu). Ames, Iowa is a progressive community of 60,000, located approximately 30 minutes north of Des Moines, and recently voted second best most livable small city in the nation (see www.amescvb.com).

Iowa State University is an equal opportunity employer committed to excellence through diversity and strongly encourages applications from all qualified applicants, including women, under-represented minorities, and veterans. ISU is highly responsive to the needs of dual career couples, has policies fostering work-life balance, and is an NSF ADVANCE institution. All faculty members are expected to exhibit and convey good citizenship and collegiality within the program, the department, the college, and the university, and maintain the highest standards of integrity and ethical behavior.

For more information or to apply, please use this link:

<http://www.iastatejobs.com/postings/22159>.

King Abdullah University of Science and Technology (KAUST)

Professor (all levels) in Data Integration

King Abdullah University of Science and Technology (KAUST) (<http://www.kaust.edu.sa>) is seeking a highly motivated and skilled faculty member for the Bioinformatics track whose research focuses on data integration.

KAUST is an international, graduate-level research university dedicated to advancing science and technology through interdisciplinary research, education, and innovation. Located on the shores of the Red Sea in Saudi Arabia, KAUST offers superb research facilities, generous assured research funding, and internationally competitive salaries, attracting top international faculty, scientists, engineers, and students to conduct fundamental and goal-oriented research to address the world's pressing scientific and technological challenges in the areas of food, water, energy, and the environment.

The successful applicant is expected to develop world-leading research in domain of data integration with focus on novel approaches for efficient and accurate targeted information extraction. The faculty member will be part of the Computational Bioscience Research Center (CBRC) within the Computer, Electrical and Mathematical

Sciences and Engineering (CEMSE) Division. The position will remain open until filled.

Requirements:

PhD or equivalent in a Computer Science, Mathematics or Engineering discipline. Candidates should be well-established within the research field relevant to the position grade. They should demonstrate original research and experience at the highest international level.

Responsibilities and tasks:

Research competence in the following areas is preferred:

- New computational methods for analysis of and combining/integrating data and information from disparate existing biological repositories/databases/sources including those containing experimental data.
- Analysis of biological networks.

Visit cemse.kaust.edu.sa to apply.

Applications will be considered until the positions are filled but not later than April 15, 2017.

Prospective candidates are advised to apply as soon as possible.

King Abdullah University of Science and Technology (KAUST)

Faculty Positions in Visual Computing

The Computer, Electrical, and Mathematical Sciences and Engineering Division at King Abdullah University of Science and Technology (KAUST) invites applications for faculty positions in visual computing. We particularly encourage applications by female candidates as well as applications for junior positions (Assistant professor rank), although we will consider outstanding candidates of any demographic.

KAUST is seeking candidates with an established track record of research in one of the subareas of visual computing, with visualization being a topic of particular interest. Successful candidates will have a PhD in Computer Science or related fields, as well as a strong publication record in top-tier conferences and journals. Senior candidates must have demonstrated strong leadership in the field. Successful candidates will be appointed within the Computer Science program, and are expected to engage with the KAUST Visual Computing Center (VCC).

The VCC is KAUST's hub for research activities spanning all areas of visual computing, ranging including imaging, computer vision, computer graphics, and visualization, as well as interdisciplinary applications of visual computing. The VCC offers a unique combination of an intellectually stimulating environment and access to superb facilities, including large-scale virtual reality installations in the KAUST Visualization Core Lab (KVL) and KAUST's 5 Petaflop/s Shaheen-2 supercomputer.

KAUST is an international, graduate research university dedicated to advancing science and technology through interdisciplinary research, education, and innovation. Located on the shores of the Red Sea in Saudi Arabia, KAUST offers superb research facilities, generous assured research funding, and internationally competitive salaries, attracting top international faculty, scientists, engineers, and students to conduct

curiosity-driven and goal-oriented research to address the world's pressing scientific and technological challenges related sustainability in energy, water, food, and the environment.

Please apply via the <http://cemse.kaust.edu.sa> employment site. Please include the names of three references for Assistant Professor positions and at least six for senior positions. Applications will be considered until the positions are filled but not later than April 15, 2017. Prospective candidates are advised to apply as soon as possible.

King Abdullah University of Science and Technology (KAUST)

Professor (all levels) in Bioinformatics and Computational Biology

King Abdullah University of Science and Technology (KAUST) (<http://www.kaust.edu.sa>) is seeking a highly motivated and skilled faculty member for the Bioinformatics track whose research focuses on development of methods and tools for Bioinformatics and Computational Biology.

KAUST is an international, graduate-level research university dedicated to advancing science and technology through interdisciplinary research, education, and innovation. Located on the shores of the Red Sea in Saudi Arabia, KAUST offers superb research facilities, generous assured research funding, and internationally competitive salaries, attracting top international faculty, scientists, engineers, and students to conduct fundamental and goal-oriented research to address the world's pressing scientific and technological challenges in the areas of food, water, energy, and the environment.

The successful applicant is expected to develop world-leading research in domain of bioinformatics/computational biology with focus on development of novel computational approaches for efficient and accurate methods of analyzing biological phenomena at molecular level. The faculty member will be part of the Computational Bioscience Research Center (CBRC) within the Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division. The position will remain open until filled.

Requirements:

PhD or equivalent in a Computer Science, Mathematics or Engineering discipline. Candidates should be well-established within the research field relevant to the position grade. They should demonstrate original research and experience at the highest international level.

Responsibilities and tasks:

Research competence in the following areas is preferred:

- Analysis of next generation sequencing (NGS) and other 'omics' data (e.g. CAGE, ChIP-Seq, DHS, RNA-Seq, Ribo-Seq, proteomic, metabolic and NMR spectra, etc.).
- Signaling, regulatory and metabolic pathways analysis.
- Development of tools (web-based and stand-alone) suited for efficient computational biology/bioinformatics.

Visit cemse.kaust.edu.sa to apply.

Applications will be considered until the positions are filled but not later than April 15, 2017.

Prospective candidates are advised to apply as soon as possible.

Marist College

The Marist College Department of Computing Technology currently seeks applications for two tenure-track positions in Computer Science and Information Technology & Systems.

We welcome candidates who have the following teaching and research interests: software development, systems analysis and design, data management and information security, data sciences and analytics, and cloud computing and networks. Applicants with the ability to teach in multiple areas across several disciplines will receive preference. Applicants must be willing to teach undergraduate and graduate courses in both traditional on-ground and on-line environments. Required duties outside the classroom include scholarly activities that result in peer-reviewed publications as well as engagement in college and professional services such as advising/mentoring students, serving on department, school, or college committees.

Candidates must have a doctoral degree in Computer Science, Information Systems, or a closely related field. We will consider ABDs in an appropriate field who will complete their dissertation within one year of hire. Evidence of excellence in teaching and scholarly work is required. Excellent written and oral communication skills are required. Industry and/or consulting experience is highly desirable. As our programs host a diverse population, the proven ability to work effectively in a multicultural environment is highly regarded.

Marist College is an independent and comprehensive liberal arts institution located in New York's historic Hudson River Valley. Situated on 210 acres overlooking the Hudson River, it enrolls 4,791 traditional undergraduate, 898 full and part-time graduate and 460 continuing education students. Marist also has a branch campus in Florence, Italy, and extension sites throughout New York. Marist has been recognized for excellence by *U.S. News & World Report*, *TIME Magazine*, *The Princeton Review's Best 380 Colleges*, and *Baron's Best Buys in College Education* and is noted for being a pioneer in the area of online degree programs.

To learn more or to apply, please visit <http://jobs.marist.edu>. Only online applications are accepted.

Marist College is strongly committed to the principle of diversity and is especially interested in receiving applications from members of ethnic and racial minority groups, women, individuals with disabilities, veterans, and persons from other under-represented groups.

**AN EQUAL OPPORTUNITY/AFFIRMATIVE
ACTION EMPLOYER**

Max Planck Institute

The Max Planck Institute for Informatics, as the coordinator of the Max Planck Center for Visual Computing and Communication (MPC-VCC), invites applications for

Junior Research Group Leaders in the Max Planck Center for Visual Computing and Communication

The Max Planck Center for Visual Computing and Communications offers young scientists in information technology the opportunity to develop their own research program addressing important problems in areas such as

- image communication
- computer graphics
- geometric computing
- imaging systems
- computer vision
- human machine interface
- distributed multimedia architectures
- multimedia networking
- visual media security

The center includes an outstanding group of faculty members at Stanford's Computer Science and Electrical Engineering Departments, the Max Planck Institute for Informatics, and Saarland University.

The program begins with a preparatory 1-2 year postdoc phase (**Phase P**) at the Max Planck Institute for Informatics, followed by a two-year appointment at Stanford University (**Phase I**) as a visiting assistant professor, and then a position at the Max Planck Institute for Informatics as a junior research group leader (**Phase II**). However, the program can be entered flexibly at each phase, commensurate with the experience of the applicant.

Applicants to the program must have completed an outstanding PhD. Exact duration of the preparatory postdoc phase is flexible, but we typically expect this to be about 1-2 years. Applicants who completed their PhD in Germany may enter Phase I of the program directly. Applicants for Phase II are expected to have completed a postdoc stay abroad and must have demonstrated their outstanding research potential and ability to successfully lead a research group.

Reviewing of applications will commence on **01 Jan 2017**. The final deadline is **31 Jan 2017**. Applicants should submit their CV, copies of their school and university reports, list of publications, reprints of five selected publications, names of 3-5 references, a brief description of their previous research and a detailed description of the proposed research project (including possible opportunities for collaboration with existing research groups at Saarbrücken and Stanford) to:

Prof. Dr. Hans-Peter Seidel
Max Planck Institute for Informatics,
Campus E 1 4, 66123 Saarbrücken, Germany;
Email: mpc-vcc@mpi-inf.mpg.de

The Max Planck Center is an equal opportunity employer and women are encouraged to apply.

Additional information is available on the website www.mpc-vcc.org.

National Taiwan University

The Department of Computer Science and Information Engineering, and the Graduate Institute of Networking and Multimedia at National Taiwan Univ. have faculty openings at all ranks beginning in August 2017. Highly qualified candidates in all areas of computer science are invited to apply. A Ph.D. or its equivalent is required. Applicants are expected to conduct outstanding research and be committed to teaching. Candidates should check <http://www.csie.ntu.edu.tw/>

faculty_recruiting/ for submitting applications. The deadline is February 15, 2017. Contact Prof. Chih-Jen Lin at faculty_search@csie.ntu.edu.tw for any questions. An early submission is strongly encouraged.

North Dakota State University

Assistant/Associate Professor

Computer Science Department seeks to fill a tenure-track

Assistant/Associate Professor position in Computational Biology, Bioinformatics, Big Data, Data Mining, or a related area starting Fall, 2017 or thereafter. This individual would become a member of an existing multi-disciplinary research group investigating cancer detection and treatment.

PhD required. NDSU offers degrees at all levels in Computer Science and Software Engineering. Research and teaching excellence is expected, normal teaching load is three courses/year. The department has 16 Faculty; 4 Lecturers; 200 graduate students (Master's and Ph.D) and 400 BS/BA undergraduate majors.

NDSU is a top national research university. Fargo is a clean, growing metropolitan area of 250,000 that consistently ranks near the top in national quality-of-life surveys. We have low levels of crime and pollution, excellent schools, short commutes, and proximity to the

Minnesota lake country. There is a symphony, an opera, a domed stadium, a community theater, three colleges, a research technology park and many other amenities.

See <https://jobs.ndsu.edu/> for more information.

NDSU is an EEO/AA-MF/Vet/Disability

Job closing will be March 1, 2017 or until the position is filled

Oakland University

Department of Computer Science and Engineering

Tenure-track Faculty Positions

The Department of Computer Science and Engineering (<http://www.cse.secs.oakland.edu>) at Oakland University seeks applications for two tenure-track assistant professor positions, starting on August 15, 2017. Applicants must have completed a Ph.D. in Computer Science, Information Technology, or a closely related field by the appointment date. Candidates must demonstrate success in research and a strong commitment to excellence in teaching. While excellent candidates from all areas of computer science will be considered, candidates in Data Science, Operating Systems, Database, Human Computer Interface, and Gaming are especially encouraged to apply.

Evaluation of applications will start on *Jan 5th 2017* and continue until the positions are filled. Applicants should upload a letter of intent, a statement of research, a statement of teaching, curriculum vitae (cv), unofficial transcripts, and the names and contact information for three references to <http://jobs.oakland.edu/postings/9292>.

The Department of Computer Science and Engineering at Oakland University is currently offering B.S. degrees in Computer Science and in

Information Technology, M.S. degrees in Computer Science, Cyber Security, and Software Engineering and Information Technology, and a Ph.D. degree in Computer Science and Informatics.

Oakland University is an ADVANCE institution, one of a limited number of universities in receipt of NSF funds in support of our commitment to increase diversity and the participation and advancement of women and underrepresented minorities in the STEM fields. Oakland University is an Affirmative Action/Equal Opportunity Employer. For more information about the Oakland University, please visit <http://www.oakland.edu>.

Portland State University

The Computer Science Department at Portland State University (PSU) invites applications for a tenure-track faculty position at the assistant professor level, to begin Fall 2017. Exceptional applicants at other ranks will also be considered.

The department currently has twenty-two tenure-track faculty members, including three NSF CAREER Award winners and two ACM Fellows. The department offers an ABET-accredited B.S., both a thesis and a non-thesis M.S., and a Ph.D. in Computer Science. The department currently serves approximately 800 undergraduates and 130 graduate students. Our teaching loads give faculty time to maintain funded research programs. Further information about the department is available at <http://www.pdx.edu/computer-science>.

PSU is the largest urban university in Oregon and is known nationally for its community engagement and sustainability initiatives. Its campus in downtown Portland is well served by public transit and offers proximity to world-class restaurants, cultural venues and outdoor activities. PSU's urban setting provides a living laboratory for research and easy access to collaborations in industry, academia and government. Current local collaborations include Intel, Oregon Health & Science University, and Oregon Department of Transportation. Portland is the home of a burgeoning software industry, including Puppet Labs, Urban Airship, Elemental Technologies, Janrain, and Webtrends.

Queens College

The Department of Computer Science at Queens College of CUNY is accepting applications for a tenure-track position in Cybersecurity at the Assistant Professor level starting Fall 2017. Consult <http://www.cs.qc.cuny.edu> for further information.

Rutgers University

The Computer Science Department at Rutgers University invites applications for several tenure-track Assistant Professor positions focusing on (a) Data Science and AI, and (b) Distributed Networks and Systems. Responsibilities include teaching undergraduate and graduate level courses in various fields of Computer Science and supervision of PhD students based on funded projects. The appointments will start September 2017.

Qualifications: Applicants should show evidence of exceptional research promise with potential for external funding, and commitment to quality advising and teaching. Hired candidates must complete their Ph.D. in Computer Science or a closely related field by August 31, 2017. Applications received by January 13, 2017 will be given priority.

To apply for the Data Science and AI positions, go to: apply.interfolio.com/39330.

To apply for the Distributed Networks and Systems position, go to: apply.interfolio.com/39389.

If you have further questions, please email the hiring committee: kagarwal@cs.rutgers.edu.

State University of New York at Binghamton

Department of Computer Science
<http://cs.binghamton.edu>

The Computer Science Department at Binghamton University has three tenure-track assistant professor positions beginning Fall 2017. Applicants should have a Ph.D. in Computer Science or related discipline, a strong research record, and a commitment to research and teaching. Qualified applications are invited from candidates with specializations in any of the following areas: (1) healthcare systems, bioinformatics, or computational biology, (2) operating systems or embedded systems, and (3) data science, big data, or machine learning.

Further details and application information are available at: <http://binghamton.interviewexchange.com>.

Applications will be reviewed until positions are filled. First consideration will be given to applications received by **January 20, 2017**.

Binghamton University is an Equal Opportunity/ Affirmative Action/Disability/Veterans Employer.

The Ohio State University Assistant Professor of Practice, Computer Science and Engineering Columbus, OH

The Computer Science and Engineering Department at The Ohio State University seeks to fill one Assistant Professor of Practice clinical-track position starting in Autumn 2017. Highly qualified applicants at more senior levels will also be considered. This is a full-time, non-tenure-track three-year faculty position which is renewable.

Clinical-track faculty members in the Department will develop, enhance, and teach courses across the computer science curriculum, in particular those emphasizing professional computing practices and practical design and implementation projects such as junior project courses and capstone design courses; and work on innovative projects, as part of independent studies with individual students, at the undergraduate and/or MS level. Clinical faculty may also collaborate on and/or lead research and development projects, including collaborations with industry, and large-scale research and development projects, as well as education-oriented projects.

Required: Applicants must have a Ph.D. in Computer Science or closely related field, or equivalent professional experience. Additional

details on expected qualifications are available at cse.osu.edu/departments/faculty-recruiting.

To apply, please arrange for a CV and three letters of recommendation to be sent by email to fsearch@cse.ohio-state.edu. Review of applications will begin on January 1, 2017 and will continue until the position is filled.

The Ohio State University is an equal opportunity employer. All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation or gender identity, national origin, disability status, or protected veteran status.

University of South Carolina College of Engineering and Computing Department of Computer Science and Engineering Faculty Position in Cybersecurity

The University of South Carolina invites applications for a tenure-track faculty position at the rank of assistant professor in the Department of Computer Science and Engineering, starting Fall 2017. The Department will consider exceptional candidates in any cybersecurity areas, but is particularly interested in candidates whose primary research expertise is in design and development of secure software architecture, development of secure and reliable applications, software vulnerability testing, or reverse engineering.

Eligible candidates should possess a Ph.D. degree in computer science, software engineering, or a closely related field and a demonstrated record of research accomplishments in the area of cybersecurity. Prior teaching experience is preferable but not required.

Successful candidates will be expected to develop internationally recognized, externally funded research programs that complement existing strengths in the Department. Applicants will also be encouraged to participate in interdisciplinary projects. To serve the diverse student population of the university, all faculty members are expected to engage and mentor students from underrepresented groups.

Application review begins December 1st, 2016 and remains open until the position is filled. Interested applicants should send one complete PDF file that includes a cover letter, curriculum vitae, a concise description of research and teaching plans, and names and contact information of three references to search@cse.sc.edu.

Founded in 1801 and one of the three public universities in South Carolina, the University of South Carolina (USC) is located in Columbia, the capital and technology center of South Carolina. USC is the flagship university of the state with a diverse student population of 33,772 students. USC is one of only 32 public universities to earn the Carnegie Foundation's top-tier designations in research activity and community engagement.

USC is designated by the National Security Agency and the Department of Homeland Security as a National Center of Academic Excellence in Information Assurance and Cyber Defense Education and Research. Cybersecurity education and research activities are centered in the Department of Computer Science and Engineering in the College of Engineering and Computing. The Department offers B.S. degrees in Computer Science, Computer Information

Systems, and Computer Engineering, M.S. and Ph.D. degrees in Computer Science and Computer Engineering, an M.S. in Software Engineering and a Graduate Certificate in Cyber Security Studies. The Department has 21 full-time faculty members (10 of whom are NSF CAREER award recipients), an undergraduate enrollment of 872 students, a graduate enrollment of 175 students, and over \$1.5 million in annual research expenditures.

The University of South Carolina is an Affirmative Action/Equal Opportunity Employer. Minorities and women are especially encouraged to apply. The University of South Carolina does not discriminate in educational or employment opportunities or decisions for qualified persons on the basis of race, color, religion, sex, national origin, age, disability, sexual orientation or veteran status.

University of California, Santa Barbara

The Department of Computer Science (CS) and the College of Creative Studies (CCS) at the University of California, Santa Barbara seek applications for a joint faculty Lecturer position with Potential Security of Employment (LPSOE, similar to tenure-track), with a start date of Fall Quarter 2017.

At the University of California, LPSOE positions lead to security of employment (SOE, similar to tenure), and are faculty positions designed to meet the long-term instructional needs of the University. This position is intended for an innovative individual with an enthusiasm for, and breadth of knowledge in, Computer Science and its emerging applications in diverse fields. A successful applicant should be committed to student mentoring, to improving undergraduate computer science education and diversity, and to recruiting top high-school graduates, in collaboration with other faculty.

The teaching load will be evenly divided between the CS department and the CCS Computing program. The Department of Computer Science at UCSB is housed in the College of Engineering, has around 35 faculty, 400 undergraduates and 175 graduate students, and offers a traditional BS degree in Computer Science. The College of Creative Studies (CCS) Computing Program is a small, tight-knit community of around 35 talented, passionate, self-directed learners in the context of a major research university, with a focus on early involvement in undergraduate research in Computer Science, and related creative activity. Candidates are urged to learn more about the UCSB Department of Computer Science and the distinctive nature of the College of Creative Studies at their respective websites: <http://www.cs.ucsb.edu> and <http://www.ccs.ucsb.edu>.

Preferred qualifications: A demonstrated record of teaching excellence.

Minimum qualifications: A Ph.D. in Computer Science or a related field is required.

The position is to begin with the 2017-18 academic year, and salary will be commensurate with experience. More detail about the position, and how to apply is available at the UCSB AP Recruit website: <https://recruit.ap.ucsb.edu/apply/JPF00882>.

Completed applications received by Dec 12, 2016 will be given primary consideration, thereafter, the position remains open until filled.

The department is especially interested in candidates who can contribute to the diversity and excellence of the academic community through research, teaching, and service.

The University of California is an Equal Opportunity/Affirmative Action Employer, and 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.

University of Akron Assistant Professor Computer Science

University of Akron Computer Science seeks tenure track Assistant Professor in security & privacy for Fall 2017.

University of Delaware Teaching Faculty Department of Computer & Information Sciences

Applications are invited for a full-time, nontenure track faculty position to begin September 1, 2017. For application and details, please visit: <https://apply.interfolio.com/39414>.

US Naval Academy

USNA Electrical & Computer Engineering is seeking applicants to fill tenure-track Assistant Professor positions in Computer Engineering. Applicants with teaching & research interests in all areas of computer engineering will be considered. Applications accepted through "Apply URL" only.

Yeshiva College

Yeshiva College of Yeshiva University invites applications for tenure-track faculty at the assistant professor level in computer science (undergraduate), with a focus on data science. Candidates must have a Ph.D. or its equivalent in professional experience.

Responsibilities of the position include:

- Teaching courses in the data science track, including Machine Learning and Modern Data Management, as well as core C.S. courses
- Ensuring that the department's course offerings and priorities remain state-of-the-art and aligned with industry requirements, expectations, and applications, in the area of data science
- Mentor/guide students in their intellectual and professional development
- Further the state of the art in data science via open source contributions and research
- Potential to teach at the graduate level in the future

The ideal candidate will have (1) demonstrated expertise in both practice and scholarship in the area of data science, especially machine

learning and related areas of modern data management, (2) proven teaching ability, and (3) industry experience.

Salary is competitive. The position is available Fall 2017. Please submit curriculum vitae and three letters of recommendation to diamant@yu.edu and vkelly@yu.edu. Please provide copies of papers and patents, and links to any open source work.

About the University

Founded in 1886, Yeshiva University (YU) has a strong tradition of combining Jewish scholarship with academic excellence and achievement in the liberal arts, sciences, medicine, law, business, social work, Jewish studies, education psychology, social work, and more.

We are a leading global educational institution that employs 2,000 people across our various campus locations -- Wilf Campus, Beren Campus, Brookdale Center, Resnick Campus in the Bronx, the Gruss Institute in Jerusalem, the Boys High School in Manhattan and the Girls High School in Queens. From the distinguished faculty who teach here, to the dedicated staff, we work to fulfill our mission: to "bring wisdom to life" through all that we teach, by all that we do and for all those we serve. We seek to attract and retain engaged and committed individuals who contribute to an exciting working environment, where there is a sense of community and belonging, balanced with a significant cross section of people from diverse backgrounds working and studying together.

Yeshiva University is an equal opportunity employer committed to hiring minorities, women, individuals with disabilities and protected veterans.

York University

The Department of Electrical Engineering and Computer Science (EECS) at York University is seeking an outstanding candidate for an alternate-stream (teaching-focused) tenure-track position at the Assistant Lecturer level to serve as course director and laboratory instructor. While outstanding candidates in all areas of EECS will be considered, we are especially interested in those with strong abilities to develop, manage and deliver laboratory sections associated with courses entailing major programming and software development components. Priority will be given to candidates eligible for licensure as Professional Engineers in Ontario. Complete applications must be received by 15 March 2017. Full job description and application details are available at: <http://lassonde.yorku.ca/new-faculty/>. All York University positions are subject to budgetary approval. York University is an Affirmative Action (AA) employer. The AA Program can be found at <http://www.yorku.ca/acadjobs> or a copy can be obtained by calling the AA office at 416-736-5713. All qualified candidates are encouraged to apply; however, Canadian citizens and permanent residents will be given priority.



DOI:10.1145/3022109

Dennis Shasha

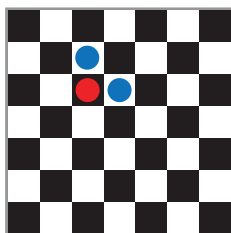
Upstart Puzzles

Open Field Tic-Tac-Toe

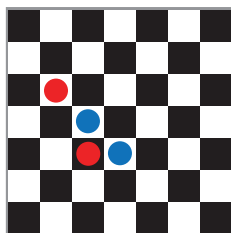
IN THE SPIRIT OF GOMOKU, two people play a version of the classic paper-and-pencil game tic-tac-toe but on an infinite checkerboard. In it, a player wins by getting four pieces in a row—vertically, horizontally, or diagonally.

Warm-up. Can the first player—blue—force a win in seven turns or less, where a turn consists of both blue and red placing pieces.

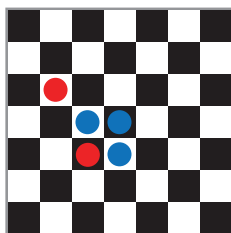
Solution to warm-up. The first player can force a win in five turns. Blue moves. No matter where red moves, blue can, in the second move, have two in a row.



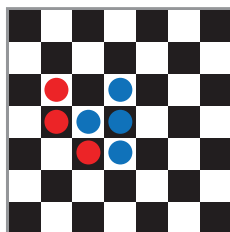
Red must now respond to prevent blue from having three in a row that is open on both ends. So R blocks, giving us something like



Blue can now force a two-by-two fork with

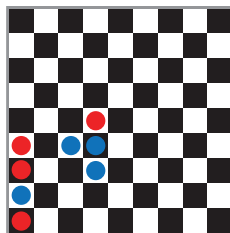


No matter where red goes, blue can force an open-ended vertical or horizontal line with three blues, as in

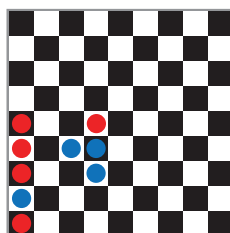


So... now that we know how it works, let us try it for some other problems.

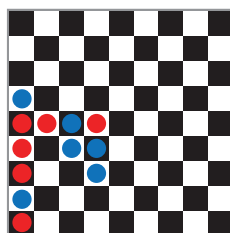
Suppose we have a board with a nine-by-nine grid with the following configuration, and red is about to take the next turn. Can either side force a win?



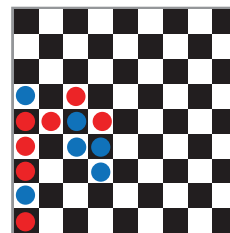
Solution. Yes, red can force a win. Red threatens...



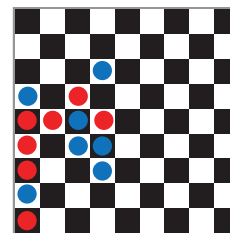
Blue then red then blue...



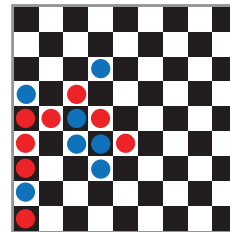
Red continues to threaten...



Blue responds...



Red now gets three in a row...



Upstart. Suppose the board is a six-by-six grid with a red exactly in every corner? Blue moves first. There is no limit on the number of turns. Can either side force a win?

All are invited to submit their solutions to upstartpuzzles@cacm.acm.org; solutions to upstarts and discussion will be posted at <http://cs.nyu.edu/cs/faculty/shasha/papers/cacmpuzzles.html>

Dennis Shasha (dennishasha@yahoo.com) is a professor of computer science in the Computer Science Department of the Courant Institute at New York University, New York, as well as the chronicler of his good friend the omniheurist Dr. Ecco.

Copyright held by the author.



The 6th ACM International Symposium on Pervasive Displays

Lugano, Switzerland, 7-9 June, 2017

Important Dates

Paper Submissions:

3 February 2017

Notifications by:

10 April 2017

Early Registration ends:

5 May 2017

The ACM International Symposium on **Pervasive Displays** is the premier venue for discussing opportunities and challenges raised by the emergence of pervasive display systems as a **new communication medium** for public and semi-public spaces. As a targeted topic venue, Pervasive Displays offers participants a unique **opportunity to network** with a diverse but focused research community, resulting in an extremely lively event with all the energy and excitement that characterizes the emergence of a **new research area**.

We are looking forward to seeing you in Lugano in June 2017!

Image Credits: "Sunset on Lugano from Monte Bre" © 2010 Francesco Rossi Photography

CC BY-NC-ND 2.0 (cropped from original Flickr image)

General Chair

Marc Langheinrich

Università della Svizzera italiana (USI), CH



Program Chair

Sarah Clinch

University of Manchester, UK



Program Committee

Florian Alt University of Munich (LMU), Germany
Matthias Baldauf Vienna University of Technology, Austria
Keith Cheverst Lancaster University, UK
Nigel Davies Lancaster University, UK
Ivan Elhart Università della Svizzera italiana (USI), Switzerland
Ava Fatah Gen. Schieck University College London, UK
Marcus Foth Queensland University of Technology, Australia
Sven Gehring German Research Center for Artificial Intelligence (DFKI), Germany
Simo Hosio University of Oulu, Finland
Rui José University of Minho, Portugal
Mohamed Khamis University of Munich (LMU), Germany
Christian Kray University of Münster, Germany
Hannu Kukka University of Oulu, Finland
Alessio Malizia Brunel University London, UK

Scott McQuire University of Melbourne, Australia
Timo Ojala University of Oulu, Finland
Florian Perteneder University of Applied Sciences Upper Austria, Austria
Aaron Quigley University of St. Andrews, UK
Roman Rädle Aarhus University, Denmark
Enrico Rukzio University of Ulm, Germany
Stacey Scott University of Waterloo, Canada
Marcos Serrano University of Toulouse (IRIT), France
Nick Taylor University of Dundee, UK
Martin Tomitsch University of Sydney, Australia
Andrew Vande Moere K.U.Leuven, Belgium
Jim Wallace University of Waterloo, Canada
Alexander Wiethoff University of Munich (LMU), Germany
Julie R. Williamson University of Glasgow, UK

Università
della
Svizzera
italiana

www.pervasivedisplays.org

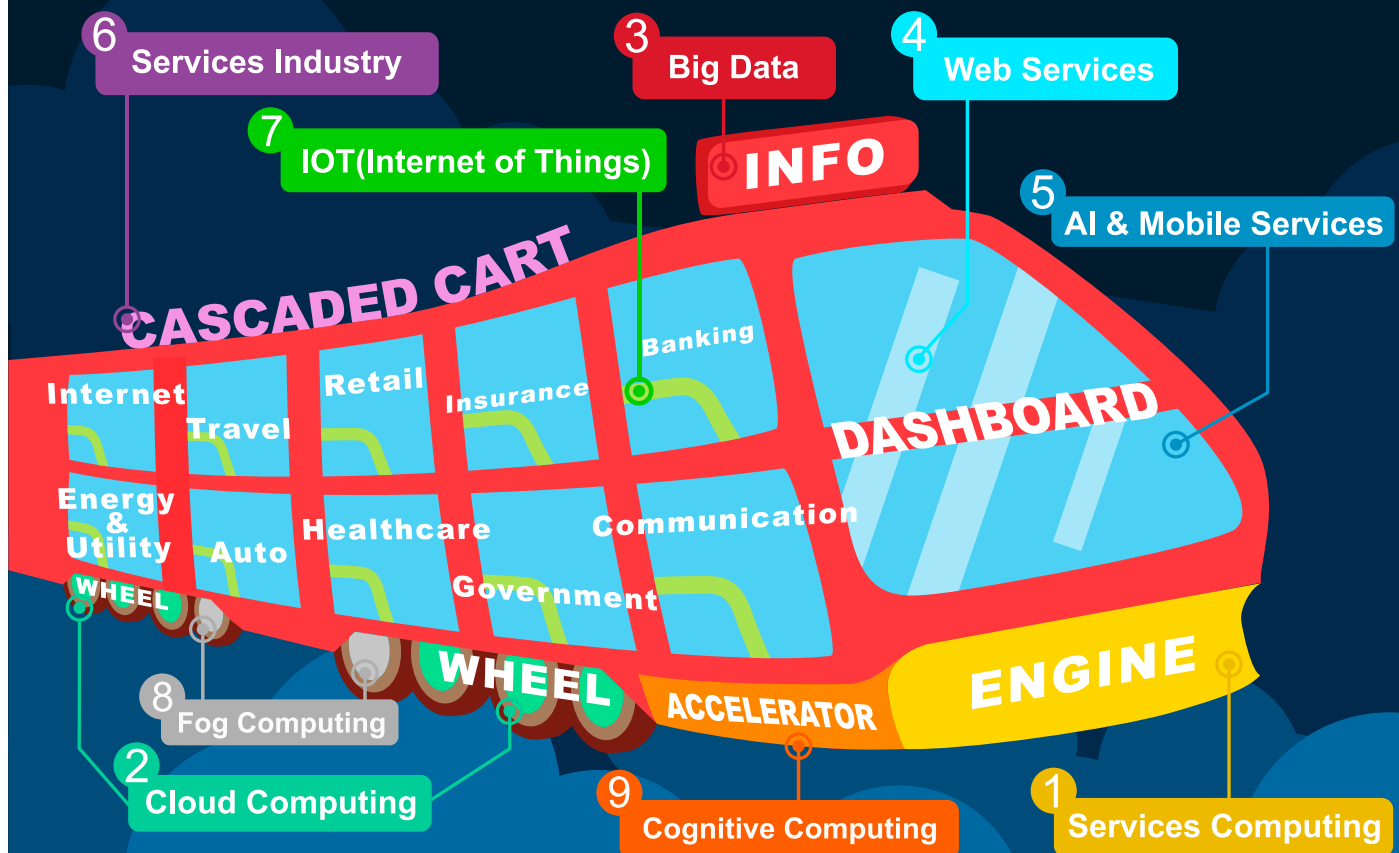


2017 IEEE Services Congress IEEE BigData Congress

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 The 2nd International Congress on Internet of Things (ICIOT 2017)*
- ⑧ The 1st International Conference on Fog Computing (ICFC 2017)*
- ⑨ The 1st International Conference on Cognitive Computing (ICCC 2017)*



Submission Deadlines

1/12/2017: ICWS 2017 (<http://icws.org>)
 1/12/2017: CLOUD 2017 (<http://theCloudComputing.org>)
 1/19/2017: SCC 2017 (<http://theSCC.org>)
 1/19/2017: AIMS 2017 (<http://theMobileServices.org>)
 1/26/2017: BigData Congress 2017 (<http://ieeeBigData.org>)
 1/26/2017: SERVICES 2017 (<http://ServicesCongress.org>)

2/1/2017: ICIOT 2017 (<http://iciot.org>)
 2/1/2017: ICFC2017 (<http://theFogComputing.org>)
 2/1/2017: ICC2017 (<http://theCognitiveComputing.org>)

Contact: confs@ServicesSociety.org



*IEEE Approval Pending