

COMMUNICATIONS

CACM.ACM.ORG

OF THE

ACM

03/2013 VOL.56 NO.03

Exploration and Mapping with Autonomous Robot Teams

Exact Exponential Algorithms

Decoding Dementia

To Boycott or Not to Boycott

Can Computing Professionals Reduce Gun Violence?

Making the Mobile Web Faster



SENSE THE TRANSFORMATION



SIGGRAPH
ASIA2013
HONG KONG

CONFERENCE 19 NOV - 22 NOV
EXHIBITION 20 NOV - 22 NOV

HONG KONG CONVENTION
AND EXHIBITION CENTRE

SA2013.SIGGRAPH.ORG

LEAD SPONSOR



SPONSORED BY





Congratulations

2012 ACM Distinguished Members

ACM honors 41 new inductees as Distinguished Members in recognition of their contributions to both the practical and theoretical aspects of computing and information technology

2012 ACM Distinguished Educators

Joel C. Adams

Calvin College

Stephen C. Cooper

Stanford University

Dan Garcia

University of California, Berkeley

Lillian (Boots) Cassel

Villanova University

Wanda P. Dann

Carnegie Mellon University

Barbara Boucher Owens

Southwestern University

2012 ACM Distinguished Engineers

Murthy Devarakonda

IBM T. J. Watson Research Center

Kenneth Russell Fast

Electric Boat Corporation

Michel Hack

IBM T. J. Watson Research Center

2012 ACM Distinguished Scientists

Nancy M. Amato

Texas A&M University

Daniel A. Jiménez

Texas A&M University

Sudipta Sengupta

Microsoft Research

Ruth Iris Bahar

Brown University

Kimberly Keeton

Hewlett-Packard Laboratories

Sandeep K. Shukla

Virginia Tech

Edward Wes Bethel

Lawrence Berkeley National Laboratory

Angelos Dennis Keromytis

Columbia University

Mei-Ling Shyu

University of Miami

Athman Bouguettaya

RMIT University, Melbourne, Australia

Latifur Khan

University of Texas at Dallas

Peter F. Sweeney

IBM T. J. Watson Research Center

Ian Brown

University of Oxford

Ninghui Li

Purdue University

Peri Tarr

IBM T. J. Watson Research Center

K. Selcuk Candan

Arizona State University

Joseph P. Loyall

Raytheon BBN Technologies

Jeffrey S. Vetter

*Oak Ridge National Laboratory and
Georgia Institute of Technology*

Naehyuck Chang

Seoul National University

Maged M. Michael

IBM T. J. Watson Research Center

Jennifer L. Welch

Texas A&M University

Chen-Nee Chuah

University of California, Davis

Michael Muller

IBM Research

Changsheng Xu

*Institute of Automation,
Chinese Academy of Sciences*

Evgeniy Gabrilovich

Google

Erich M. Nahum

IBM T. J. Watson Research Center

Torben Bach Pedersen

Aalborg University, Denmark

Franco Zambonelli

Università di Modena e Reggio Emilia

Wendy Beth Heinzelman

University of Rochester

Vijay V. Raghavan

University of Louisiana at Lafayette

Antony L. Hosking

Purdue University

Wenwu Zhu

Tsinghua University



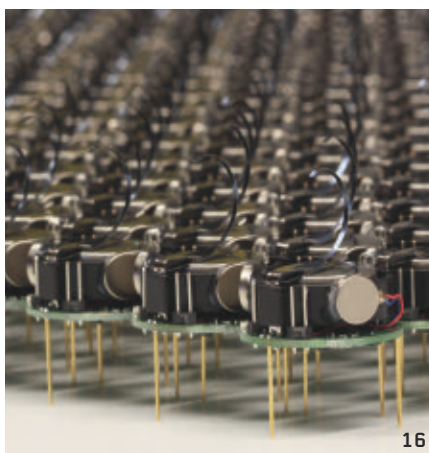
Departments

- 5 **Editor's Letter**
To Boycott or Not to Boycott
By Moshe Y. Vardi
-
- 7 **From the President**
A Revolution in India
By Vincent G. Cerf
-
- 8 **Letters to the Editor**
No Place for Old Educational Flaws in New Online Media
-
- 10 **BLOG@CACM**
Passwords Getting Painful, Computing Still Blissful
Jason Hong wonders how anyone can follow the mounting complexity of password rules, and Daniel Reed ponders the attractions of computing.
-
- 31 **Calendar**
-
- 101 **Careers**

Last Byte

- 102 **Puzzled**
Solutions and Sources
By Peter Winkler
-
- 104 **Q&A**
The Power of Distribution
Nancy Lynch talks about achieving consensus, developing algorithms, and mimicking biology in distributed systems.
By Leah Hoffmann

News

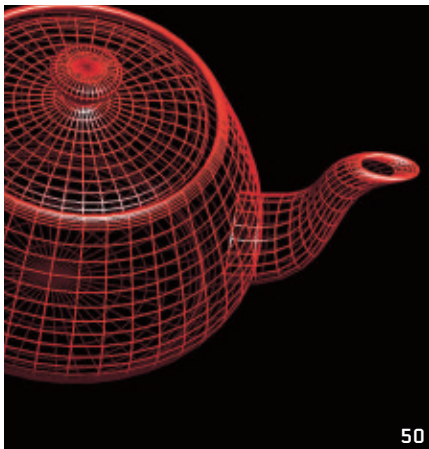


- 13 **Decoding Dementia**
Computer models may help neurologists unlock the secrets of brain disorders, from Alzheimer's to cancer.
By Neil Savage
-
- 16 **Rise of the Swarm**
Guided by collective intelligence, teams of small, simple robots could soon accomplish amazing feats.
By Gregory Mone
-
- 18 **Cybercrime: It's Serious, But Exactly How Serious?**
Symantec says \$110 billion annually while McAfee says \$1 trillion. Why can't anyone agree?
By Paul Hyman
-
- 21 **ACM Fellows Inducted**

Viewpoints

- 24 **Legally Speaking**
A Copyright Challenge to Resales of Digital Music
A currently pending case will have significant implications for secondary markets in digital goods.
By Pamela Samuelson
-
- 27 **Broadening Participation**
Academic Careers Workshop for Underrepresented Groups
A longitudinal evaluation of the application of knowledge, skills, and attitudes of ACW participants.
By Denice Ward Hood, Stafford Hood, and Dominica McBride
-
- 30 **The Profession of IT**
Moods, Wicked Problems, and Learning
Wicked problems and learning environments present tough mood challenges for leaders and teachers. Telepresence and sensory gadgets are unlikely to replace physical presence in these areas.
By Peter J. Denning
-
- 33 **Computing Ethics**
Ethics Viewpoints Efficacies
Seeking answers to ethical concerns.
By Rachelle Hollander
-
- 35 **Viewpoint**
Can Computer Professionals and Digital Technology Engineers Help Reduce Gun Violence?
Ten idea seeds.
By Jeff Johnson
-
- 38 **Viewpoint**
Funding Successful Research
A proposal for result-based funding for research projects.
By Mikkel Thorup

Practice



50

40 **Hazy: Making It Easier to Build and Maintain Big-Data Analytics**

Racing to unleash the full potential of big data with the latest statistical and machine-learning techniques.

By Arun Kumar, Feng Niu, and Christopher Ré

50 **The Story of the Teapot in DHTML**
It is easy to do amazing things, such as rendering the classic teapot in HTML and CSS.

By Brian Beckman and Erik Meijer

56 **Making the Mobile Web Faster**
Mobile performance issues? Fix the back end, not just the client.

By Kate Matsudaira



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

**About the Cover:**

A team of cooperating robots can dramatically increase the effectiveness of a human working alone. This month's cover story (p. 62) details the challenges and advantages of creating robot teams for search-and-rescue and reconnaissance tasks. Cover illustration by Coherent Images.

Contributed Articles



62

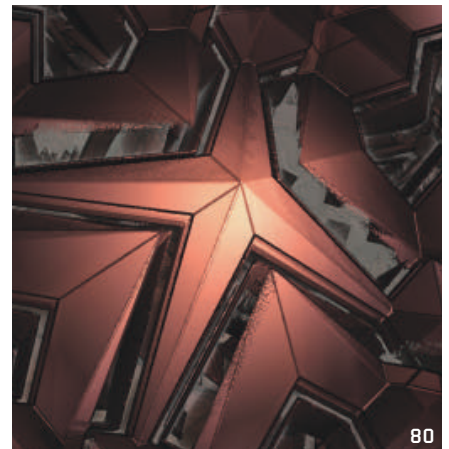
62 **Exploration and Mapping with Autonomous Robot Teams**
The MAGIC 2010 robot competition showed how well multi-robot teams can work with human teams in urban search.

By Edwin Olson, Johannes Strom, Rob Goeddel, Ryan Morton, Pradeep Ranganathan, and Andrew Richardson

71 **Mobile Social Networking Applications**
They deliver the right social service to the right user anytime, anyplace, without divulging personal data.

By Nafaâ Jabeur, Sherali Zeadally, and Biju Sayed

Review Articles



80

80 **Exact Exponential Algorithms**
Discovering surprises in the face of intractability.

By Fedor V. Fomin and Petteri Kaski

Research Highlights

90 **Technical Perspective**
Video Quality Assessment in the Age of Internet Video

By David Oran

91 **Understanding the Impact of Video Quality on User Engagement**
By Florin Dobrian, Asad Awan, Dilip Joseph, Aditya Ganjam, Jibin Zhan, Vyas Sekar, Ion Stoica, and Hui Zhang



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

John White

Deputy Executive Director and COO

Patricia Ryan

Director, Office of Information Systems

Wayne Graves

Director, Office of Financial Services

Russell Harris

Director, Office of SIG Services

Donna Cappel

Director, Office of Publications

Bernard Rous

Director, Office of Group Publishing

Scott E. Delman

ACM COUNCIL

President

Vinton G. Cerf

Vice-President

Alexander L. Wolf

Secretary/Treasurer

Vicki L. Hanson

Past President

Alain Chesnais

Chair, SGB Board

Erik Altman

Co-Chairs, Publications Board

Ronald Boisvert and Jack Davidson

Members-at-Large

Eric Allman; Ricardo Baeza-Yates;

Radia Perlman; Mary Lou Soffa;

Eugene Spafford

SGB Council Representatives

Brent Hailpern; Joseph Konstan;

Andrew Sears

BOARD CHAIRS

Education Board

Andrew McGettrick

Practitioners Board

Stephen Bourne

REGIONAL COUNCIL CHAIRS

ACM Europe Council

Fabrizio Gagliardi

ACM India Council

Anand S. Deshpande, PJ Narayanan

ACM China Council

Jianguang Sun

PUBLICATIONS BOARD

Co-Chairs

Ronald F. Boisvert; Jack Davidson

Board Members

Marie-Paule Cani; Nikil Dutt; Carol Hutchins;

Joseph A. Konstan; Ee-Peng Lim;

Catherine McGeoch; M. Tamer Ozsu;

Vincent Shen; Mary Lou Soffa

ACM U.S. Public Policy Office

Cameron Wilson, Director

1828 L Street, N.W., Suite 800

Washington, DC 20036 USA

T (202) 659-9711; F (202) 667-1066

Computer Science Teachers Association

Chris Stephenson,

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 GROUP PUBLISHING

Scott E. Delman

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

Editorial Assistant

Zarina Strakhan

Rights and Permissions

Deborah Cotton

Art Director

Andrij Borys

Associate Art Director

Margaret Gray

Assistant Art Directors

Mia Angelica Balaquiot

Brian Greenberg

Production Manager

Lynn D'Addesio

Director of Media Sales

Jennifer Ruzicka

Public Relations Coordinator

Virginia Gold

Publications Assistant

Emily Williams

Columnists

Alok Aggarwal; Phillip G. Armour;

Martin Campbell-Kelly;

Michael Cusumano; Peter J. Denning;

Shane Greenstein; Mark Guzdial;

Peter Harsha; Leah Hoffmann;

Mari Sako; Pamela Samuelson;

Gene Spafford; Cameron Wilson

CONTACT POINTS

Copyright permission

permissions@cacm.acm.org

Calendar items

calendar@cacm.acm.org

Change of address

acmhlp@cacm.org

Letters to the Editor

letters@cacm.acm.org

WEBSITE

<http://cacm.acm.org>

AUTHOR GUIDELINES

<http://cacm.acm.org/guidelines>

ACM ADVERTISING DEPARTMENT

2 Penn Plaza, Suite 701, New York, NY

10121-0701

T (212) 626-0686

F (212) 869-0481

Director of Media Sales

Jennifer Ruzicka

jen.ruzicka@hq.acm.org

Media Kit acmmidasales@cacm.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

vic@cacm.acm.org

NEWS

Co-Chairs

Marc Najork and Prabhakar Raghavan

Board Members

Hsiao-Wuen Hon; Mei Kobayashi;

William Pulleyblank; Rajeev Rastogi

VIEWPOINTS

Co-Chairs

Susanne E. Hambrusch; John Leslie King;

J Strother Moore

Board Members

William Aspray; Stefan Bechtold; Judith

Bishop; Stuart I. Feldman;

Peter Freeman; Seymour Goodman;

Mark Guzdial; Richard Heeks;

Rachelle Hollander; Richard Ladner;

Susan Landau; Carlos Jose Pereira de Lucena;

Beng Chin Ooi; Loren Terveen;

Jeannette Wing

PRACTICE

Chair

Stephen Bourne

Board Members

Eric Allman; Charles Beeler; Bryan Cantrill;

Terry Coatta; Stuart Feldman; Benjamin Fried;

Pat Hanrahan; Tom Limoncelli;

Marshall Kirk McKusick; Erik Meijer;

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

Al Aho and Georg Gottlob

Board Members

William Aiello; Robert Austin; Elisa Bertino;

Gilles Brassard; Kim Bruce; Alan Bundy;

Peter Buneman; Erran Carmel;

Andrew Chien; Peter Druschel; Carlo Ghezzi;

Carl Gutwin; James Larus; Igor Markov;

Gail C. Murphy; Shree Nayar; Bernhard

Nebel; Lionel M. Ni; Sriram Rajamani;

Marie-Christine Rousset; Avi Rubin;

Krishan Sabnani; Fred B. Schneider;

Abigail Sellen; Ron Shamir; Yoav Shoham;

Marc Snir; Larry Snyder; Manuela Veloso;

Michael Vitale; Wolfgang Wahlster;

Hannes Werthner; Andy Chi-Chih Yao

RESEARCH HIGHLIGHTS

Co-Chairs

Stuart J. Russell and Gregory Morrisett

Board Members

Martin Abadi; Sanjeev Arora; Dan Boneh;

Andrei Broder; Stuart K. Card; Jon Crowcroft;

Alon Halevy; Monika Henzinger;

Maurice Herlihy; Norm Jouppi;

Andrew B. Kahng; Xavier Leroy;

Mendel Rosenblum; David Salesin;

Guy Steele, Jr.; David Wagner;

Alexander L. Wolf; Margaret H. Wright

WEB

Chair

James Landay

Board Members

Gene Golovchinsky; Marti Hearst;

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



ACM Copyright Notice

Copyright © 2013 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@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 \$100.

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@cacm.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



Association for
Computing Machinery



Printed in the U.S.A.



DOI:10.1145/2428556.2428557

Moshe Y. Vardi

To Boycott or Not to Boycott

THERE HAS BEEN sound and fury in the Open Access movement over the past year. In December 2011, The Research Works Act (RWA) was introduced in the U.S. House of Representatives. The bill contained provisions to prohibit open access mandates for federally funded research, effectively nullifying the U.S. National Institutes of Health's policy that requires taxpayer-funded research to be freely accessible online. Many scholarly publishers, including the Association of American Publishers (AAP), expressed support for the bill. (ACM expressed objections to the bill.)

The reaction to the bill and its support by scholarly publishers has been one of sheer outrage, with headlines such as "Academic Publishers Have Become the Enemies of Science." On January 21, 2012, renowned British mathematician Timothy Gowers declared a boycott on Elsevier, a major scholarly publisher, pledging to refrain from submitting articles to Elsevier journals, as well as from serving as an editor or reviewer. The boycott movement then took off, with over 13,000 scholars having joined so far.

Frankly, I do not understand why Elsevier is practically the sole target of the recent wrath directed at scholarly publishers. Elsevier is no worse than most other for-profit publishers, just bigger, I believe. Why boycott Elsevier and not Springer, for example? The argument made by some that "we must start somewhere" strikes me as plainly unfair and unjust.

Beyond the question of whom to target with a boycott, there is the question of the morality of the boycott. Of course, authors can choose their publication venues. Also, as a scholar, I can choose which publications I am willing to support by becoming an editor, but the boycott petition also asks signato-

ries to refrain from refereeing articles submitting to Elsevier journals. This means that if you sign this petition then, in effect, you are boycotting your colleagues who have disagreed with you and chose to submit their articles to an Elsevier journal.

I believe in keeping science separate from politics. If it is legitimate to boycott publishing politics—the issue of open access is, after all, a political issue—why is it not legitimate to boycott for other political considerations? Is it legitimate to refrain from refereeing articles written by authors from countries with objectionable government behavior? Where do you draw the line to avoid politicizing science?

My perspective is that what really propelled the Open Access movement was the continuing escalation of the price of scholarly publications during the 1990s and 2000s, a period during which technology drove down the cost of scientific publishing. This price escalation has been driven by for-profit publishers. In the distant past, our field had several small- and medium-sized for-profit publishers. There was a sense of informal partnership between the scientific community and these publishers. That was then. Today, there is a small number of large and dominant for-profit publishers in computing research. These publishers are thoroughly corporatized. They are businesses with a clear mission of maximizing the return on investment to their owners

and shareholders. At the same time, the scientific community, whose goal is to maximize dissemination, continues to behave as if a partnership exists with for-profit publishers, providing them with content and editorial services essentially gratis. This is a highly anomalous arrangement, in my opinion. Why should for-profit corporations receive products and labor essentially for free?

Beyond the moral issue I raised earlier regarding the boycott, there is a more practical issue. For-profit publishers play a key role in computing-research publishing. As an example, approximately 45,000 journal articles were published in 2011 in computing research. In that same year, ACM published fewer than 1,000 journal articles, and IEEE-Computer Society published fewer than 3,500 articles. There is a small number of other non-profit publishers, but for-profit publishers produce the lion's share of computing-research journal articles. Boycotting all of them is simply not a practical option.

I do not believe, therefore, that boycotting is the right approach to the current scholarly publishing controversies. If we want to drive for-profit publishers out of business, we have to do it the old-fashioned way, by out-publishing them. If professional associations in computing research would expand their publishing activities considerably, they should be able to attract the bulk of computing articles. ACM is only a minor player in journal publishing. Why is ACM publishing fewer than 1,000 journal articles per year rather than, say, 5,000 articles? Even if this will not drive the for-profit publishers out of the computing-research publishing business, the competition would pressure them to reform their business practices, which is, after all, what we should be after.

I believe in keeping science separate from politics.

Moshe Y. Vardi, EDITOR-IN-CHIEF

interactions

EXPERIENCES | PEOPLE | TECHNOLOGY



interactions' website interactions.acm.org, is designed to capture the influential voice of its print component in covering the fields that envelop the study of people and computers.

The site offers a rich history of the conversations, collaborations, and discoveries from issues past, present, and future.

Check out the current issue, follow our bloggers, look up a past prototype, or discuss an upcoming trend in the communities of design and human-computer interaction.

FEATURES

BLOGS

FORUMS

DOWNLOADS

interactions.acm.org

Association for
Computing Machinery





Vinton G. Cerf

DOI:10.1145/2428556.2428558

A Revolution in India

I recently had the pleasure of visiting our colleagues in India, specifically in Chennai, at the ACM-India Council^a meeting. The ACM-India Council joins the ACM-China

and ACM-Europe Councils, as well as the general ACM Council, as the foundation for the internationalization efforts of our association. In addition to the ACM-India Council meeting, I was able to attend an ACM-W India event and a computer science research symposium hosted at the Indian Institute of Technology (IIT) Chennai campus.

The 2012 ACM-India elections seated P.J. Narayanan as the chair of the ACM-India Council. “PJN,” as he is known, is the Dean of Research and Development at the International Institute of Information Technology (IIIT) at Hyderabad,^b an institution set up as a public/private partnership with the state of Andhra Pradesh. His enthusiasm and skilled leadership bode well for ACM in India. The vigor of the ACM program in India is underscored by 63 student chapters and 13 professional chapters, in addition to very active participation in the ACM-W India program. At the Chennai meetings, I counted over 150 women in attendance at the ACM-W event, as well as a score or more men, including me, who were pleased by the dramatic reversal of the usual gender balance!

It is well known that India has embraced and invested in information technology (IT) as an enhancer of job creation, notably outsourcing software and service-related businesses, facilitated in part by access to the global Internet. I learned the strong program of education in computer science, engineering, and programming had its ori-

gins in the policies of Rajiv Gandhi, the youngest prime minister in India’s history, and the son of Indira Gandhi—the first and, so far, only woman to ascend to that position. The noted entrepreneur Sam Pitroda was a key advisor to Rajiv Gandhi and strongly encouraged his initiatives in the IT and telecommunications spaces. Pitroda continues his advocacy in these efforts, serving as advisor to the present Prime Minister, Manmohan Singh, and other officials of the Indian government. Among many other posts, Pitroda is also chair of the National Innovation Council.

During my visit, I had the privilege of meeting at length with Rahul Gandhi, the son of Rajiv and Sonia Gandhi and the grandson of Indira Gandhi. Newly elected to the vice presidency of the Congress Party of India, Rahul Gandhi is a potential candidate to become Prime Minister as early as 2014. Energetic, thoughtful, and a tireless advocate for modernizing India and its infrastructure, I found Gandhi to be very receptive to the potential for applying IT for the benefit of India’s 1.2 billion citizens and for improving the Indian economy. A massive fiber-networking program is in progress to bring high-speed networking to every village in India. Distribution to homes and businesses could be achieved with the use of wireless connectivity including 2G, 3G, and LTE as well as Wi-Fi. While the current population of Internet users is estimated at only 140 million today, there are 300 million data-capable mobiles in use and that number is bound to increase. The number of smart-

phones with full Internet capability is estimated at 24 million.

I was able to meet with several Internet users living in a village on the outskirts of New Delhi: two 12-year-old sixth graders, two 20-year-old IT school graduates, and two artisans (electrician and plastering) in their 30s. The sixth graders were no strangers to Internet-based applications on mobiles and were very adept at making use of laptops/desktops at a nearby Internet café. The two IT school graduates were looking for work that would allow them to apply their training to well-paying jobs in the New Delhi area. The two older workers were using their mobiles and beginning to use laptops to track down work, stay in touch with clients, and encourage “word of mouse” advertising.

While in Hyderabad, I drove around the “Cyberabad” industrial park area seeing the names of many prominent multinational companies including Google, Microsoft, Oracle, and IBM. Major Indian firms such as WIPRO and INFOSYS were also notably visible. I saw major infrastructure projects in progress including metro systems in Chennai and New Delhi. Visits with the Department of Telecommunications confirmed their commitment to the implementation of national scale digital wired and wireless network infrastructure. These efforts reinforce the potential for Massive Open Online Courses (MOOCs) to reach larger audiences with a growing interest in practical education that can be renewed and refreshed during long and varying careers. It seemed clear to me there is a tide in Indian affairs drawing the research and academic sector, the private sector, the government, and the general population toward a decidedly digital future. Moreover, it is especially satisfying to discover the ACM has a serious role to play in encouraging and facilitating progress in this direction.

Vinton G. Cerf, ACM PRESIDENT

a See <http://india.acm.org/> for further information.

b See <http://www.iiit.ac.in/institute/about>

No Place for Old Educational Flaws in New Online Media

MOSHE Y. VARDI identified important negative trends in his Editor's Letter "Will MOOCs Destroy Academia?" (Nov. 2012) concerning massive open online courses, saying, "If I had my wish, I would wave a magic wand and make MOOCs disappear..." But we should instead regard MOOCs as part of an early, awkward stage of a shift in education likely to produce something unrecognizable within even our own generation. Like journalism, retail sales, and many other fields, education is undergoing a sea change to something more fluid in time, space, and participation, as well as more peer-oriented. With lifelong learning increasingly critical today, institutions must aim for a vision of the future that finds ways to tap subject experts, as well as a proper business model that keeps both the institutions and the experts relevant. However, one thing the change does not involve is moving the old educational model, with all its flaws, to a new online medium.

Andy Oram, Cambridge, MA

Don't Give Up On the Turing Test

Exploring non-human intelligence—real and artificial—is fascinating. Consider novels like Arthur C. Clarke's *2001: A Space Odyssey* and stories like Isaac Asimov's *I, Robot*, as well as cinematic adaptations like *Blade Runner* based on Philip K. Dick's novel *Do Androids Dream of Electric Sheep?* The plot invariably revolves around machines with an intelligence level comparable to that of humans that communicate with humans, so not far from a Turing test. Fascinating, because deep down, we, as humans, believe we are unique in our level of cognition and ability to emote.

A credible intelligent agent must be able to relate to human perception, reasoning, communication, and life experience, including emotion.

In "Moving Beyond the Turing Test" (Dec. 2012), Robert M. French argued this is impossible, outlining a scenario only a human could truly understand, backed up with an example involving a series of instructions for manipulating one's fingers. He implied that answering a question about a particular step in the sequence is, and always will be, out of bounds for machines. His assertion (about answering out-of-bounds questions) was: "Don't try; accept that machines will not be able to answer them and move on."

I must disagree. My company, North Side Inc. (<http://www.northsideinc.com/>), pursues research and development toward endowing machines with verbal ability anchored in real-world knowledge. Work in this direction requires that we account for (and simulate) human perception, motor function, cognition, and emotion. Though still far from being able to pass the Turing Test, we are making good progress; for descriptions of our recent work on embodied intelligent agents with conversational ability, see our video at <http://www.botcolony.com> and my paper at http://lang.cs.tut.ac.jp/jap-tal2012/special_sessions/GAMNLP-12/papers/gamnlp12_submission_3.pdf.

A credible intelligent agent must be able to relate to human perception, reasoning, communication, and life experience, including emotion.

Credible high-fidelity agents with human-like behavior promise great technological and economic benefit in such fields as entertainment, mobile computing, e-commerce, and training. We have also found that an agent attempting to emulate human behavior (often failing) has a quirky, humorous side that makes it endearing. Why go for a humorless computer in a world where marketers dream of intelligent assistants connecting (emotionally) with their human owners? In 1996, Byron Reeves and Clifford Nass offered ample evidence for the theory that people tend to treat computers and other media as if they were real people in *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places* (<http://csli-publications.stanford.edu/site/1575860538.shtml>).

We must keep trying to make intelligent agents as credible and human-like as we know how. However, the premise of French's article was that it is time for the Turing Test to take a bow and leave the stage. Embodied artificial cognition is an extremely difficult (but fascinating) endeavor, and the benefits of success are enormous. It is way too early to even contemplate giving up.

Eugene Joseph, Montréal, Canada

Author's Response:

Joseph claims his robots are "making good progress" toward passing a full-blown Turing Test. This is delusional, cynical (perhaps in order to attract financing), or shows he does not fully understand how incredibly difficult it would be for a machine to actually pass a carefully constructed Turing Test. My point in the article was that intelligent robots, capable of meaningful interaction with humans, do not have to be Turing-Test indistinguishable from humans. Just ask Jimmy [North Side's robot] if Ayame [North Side's nominal adult human] can put her little finger all the way up her nose.

Robert M. French, Dijon, France

Teach, Don't Just Transmit, CS Knowledge

I was disappointed in Aman Yadav's and John T. Korb's Viewpoint "Learning to Teach Computer Science: The Need for a Methods Course" (Nov. 2012). There is no question that teaching anything well requires knowledge of the subject and proper pedagogical technique, both covered nicely. Left out, however, and worse, mischaracterized, was skills. With any human behavior, knowledge is only part of the equation, typically not the most important. Yadav and Korb omitted all discussion of skills, except for mistakenly calling pedagogical knowledge a "skill set" (second paragraph in their "Learning to Teach" section). Knowledge is not skill. Skills, or competencies, are the know-how that enables a teacher to assess what method, technique, demo, analogy, illustration, or exercise works best for which students in which circumstances. Competencies cannot be reduced to knowledge.

No amount of content or pedagogical knowledge can substitute for teaching skill. Generalizations, including empirical studies, concerning how to present topic X are great, but doing it well means crafting it to the students and the case at hand. I have, for almost 30 years, taught computer science, from freshman-level intro to computing to advanced graduate courses in software engineering and AI. I focus on the student(s) and what they need to grasp the concept or acquire the skills they need. I ask myself, where are they confused? What distinction are they missing? Where did they get a wrong idea? What do I know about them that would enable me to choose the analogy that works for them, how to say it so it connects, and how to motivate them to keep working on something they likely find difficult and confusing? How can I motivate them to engage with computer science at all? Also, how do I invent new examples when the usual ones don't work? And how do I assess whether students are getting the concept or skill I am teaching? Moreover, how do I respond to the student who says, "I'm just dumb"?

The National Science Foundation CS10K Project (http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf12527) may indeed produce

10,000 teachers by 2016 but will not have much influence on the number of teenagers with knowledge, skills, and, most important, interest in computing if it does not give those teachers the skills that make them teachers, not mere knowledge transmitters.

H. Joel Jeffrey, DeKalb, IL

Real Credit for Virtual Courses?

In the news story "In the Year of Disruptive Education" (Dec. 2012), Paul Hyman explored the challenge of how to award college credit for learning gained from free online courses offered by colleges and universities. The solution may emerge in two ways:

Credit by examination (CBE). Despite already being offered by many colleges as a way to give credit for knowledge, CBE also has a downside—that students typically pay the same amount of tuition as if they were taking the course and that some schools limit the awarding of credit to those students who complete some period of residency at the school; and

Government-sponsored course recognition. Like many states, Ohio has developed pseudo-course designations, called Career-Technical Assurance Guides, or CTAGs. A CTAG identifies the core content of individual courses commonly offered at colleges, technical schools, and secondary schools; private and public colleges in Ohio can choose to tie one of their courses to a CTAG "virtual" course, in which case students earning credit for a tagged course at one institution carry that credit to all colleges with a similar tagged course.

It may be that states or even countries will develop CBE for virtual courses, and colleges that tag their courses will award college credit regardless of how a student gains proficiency. A college willing to reduce the cost of CBE and waive residency requirements could unilaterally implement it. Governments are usually motivated more than the colleges themselves to offer CBE at the lowest cost possible.

Christine Wolfe, Lancaster, OH

Communications welcomes your opinion. To submit a Letter to the Editor, please limit yourself to 500 words or less, and send to letters@cacm.acm.org.

© 2013 ACM 0001-0782/13/03

Sentiment Analysis

Offline Management in Virtual Environments

Why Do Computer Talents Become Hackers?

Cybervictimization and Cybersecurity in China

The Problem with Hands-Free Dashboard Cellphones

Discriminative Learning with Latent Variables for Cluttered Indoor Scene Understanding

And the latest news about predicting neural connections, approximate computing, and digital history.

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:10.1145/2428556.2428560

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

Passwords Getting Painful, Computing Still Blissful

Jason Hong wonders how anyone can follow the mounting complexity of password rules, and Daniel Reed ponders the attractions of computing.



Jason Hong "Password Policies are Getting Out of Control"

[http://cacm.acm.org/blogs/blog-cacm/123889-](http://cacm.acm.org/blogs/blog-cacm/123889-password-policies-are-getting-out-of-control/fulltext)
[password-policies-are-getting-](http://cacm.acm.org/blogs/blog-cacm/123889-password-policies-are-getting-out-of-control/fulltext)
[out-of-control/fulltext](http://cacm.acm.org/blogs/blog-cacm/123889-password-policies-are-getting-out-of-control/fulltext)

Aug. 23, 2011

Something I learned a long time ago is that one person's inefficiency is someone else's bottom line. This simple observation explains a lot of the big problems we are facing worldwide. Rather than getting into a discussion of those thorny political topics, however, I want to use this observation as a starting point for discussing something that plagues us all: password policies.

In fact, I think I have found the most difficult password policy in existence today. It was a U.S. government website, of course. Here were the password policies the site had in place:

Password Rules:

- ▶ Minimum 8 characters;
- ▶ Must contain at least 1 capital letter;

- ▶ Must contain at least 1 lowercase letter;
- ▶ Must contain at least 1 number;
- ▶ Must contain at least 1 special character;
- ▶ Cannot contain consecutive characters (abc or cba);
- ▶ Cannot contain repeating characters (aa, bb, cc);
- ▶ Cannot contain the same character more than twice;
- ▶ Entered password must be different from last 10 passwords used; and
- ▶ Cannot be changed within 24 hours.

It actually took me about a dozen tries to create a password that covered all of this criteria, plus was something I had a chance of remembering. Here are examples of passwords that failed:

- ▶ My_P@\$Sw0rd (failed because of repeating characters)
- ▶ !USg0v8 (failed because too short)
- ▶ StuPidP@55 (failed because repeating characters)

I tried a few randomly generated passwords, guaranteed to be strong, which also failed some required criteria.

Of course, this password expires after 60 days (on a site that I only need to use every 90 days, no less). And when it did expire, it only took me an extra 15 minutes to figure out who to call to reset the password, plus a 13-minute hold, before my password was finally reset.

Makes one wonder how much real security is actually being offered with such measures, especially given the costs of staffing a helpdesk and the wasted time to end users of having to get their passwords reset.

Why do websites have such stringent password policies?

It all comes back to the opening statement: your inefficiency is someone else's bottom line. In many organizations, there is an individual whose role is to keep computing systems secure. They are the people who get yelled at when things go wrong and whose job is on the line. In extreme cases, it becomes fully rational behavior to keep increasing security, no matter what the cost is for end users, regardless of whether it is effective or not in practice. (Replace the words "computing systems" with "air travel" and we have a decent explanation for the challenges that TSA faces.)

A 2010 paper by Dinei Florencio and Cormac Herley, two researchers at Microsoft Research, presented an analysis of password policies of 75 different websites. They found that, almost counterintuitively, "[s]ome of the largest, highest value, and most attacked sites on the Internet such as Paypal, Amazon, and Fidelity Investments al-

low relatively weak passwords,” primarily because these websites earn revenue by having people login.

In contrast, it was government and university sites that tended to have stricter (and less usable) policies. They explained these results by arguing that “[t]he reason lies not in greater security requirements, but in greater insulation from the consequences of poor usability. Most organizations have security professionals who demand stronger policies, but only some have usability imperatives strong enough to push back. When the voices that advocate for usability are absent or weak, security measures become needlessly restrictive.”

Unfortunately, there are not a lot of ways forward here. Passwords are cheap and pervasive, and are not going away anytime soon. Forcing all members of Congress and all generals to personally experience the joy of using these websites themselves also is not realistic, even if highly desirable.

In the long term, we need more ways of getting the incentives of all stakeholders better aligned. Putting helpdesk costs and information security costs under the same budget and under the same person is a good start, as it would force people to think more about the relative costs and benefits of a security policy. Having customer satisfaction be part of the performance metrics for information security folks would also help. In the meanwhile, until usability thinking and holistic thinking become more pervasive in computer security, the rest of us will just have to keep suffering the pains of stricter password policies.



Daniel Reed
“Why We Compute”

<http://cacm.acm.org/blogs/blog-cacm/126408-why-we-compute/fulltext>
Sept. 2, 2011

Why do we, as researchers and practitioners, have this deep and abiding love of computing? Why do we compute?

Superficially, the question seems as innocuous as asking why the sky is blue or the grass is green. However, like both of those childhood questions, the simplicity belies the subtlety beneath. Just ask someone about Raleigh scattering or the quantum efficiency of

photosynthesis if you doubt that simple questions can unearth complexity.

At its most basic, computing is simply automated symbol manipulation. Indeed, the abstract Turing machine does nothing more than manipulates symbols on a strip of tape using a table of rules. More deceptively, the rules seem simpler than some board games. Though vacuously true, the description misses the point that symbol manipulation under those rules captures what we now call the Church-Turing thesis.

However, as deep and as beautiful as the notion of computability really is, I doubt it is the only reason most of us are so endlessly fascinated by this malleable thing we call computing. Rather, I suspect it is a deeper, more primal yearning, one that underlies all of science and engineering and that unites us in a common cause. It is the insatiable desire to know and understand.

Lessons from Astronomy

When I stood atop Mauna Kea, looking at the array of telescopes perched there, I was again struck by our innate curiosity. Operated by a diverse array of international partnerships and built on Mauna Kea at great expense, they are there because we care about some fundamental questions. What is the evolutionary history and future of the universe? What are dark matter and dark energy? Why is there anything at all?

Answers to these questions are not likely to address our current economic woes, improve health care, or address our environmental challenges. We care about the answers, nevertheless.

As I pondered the twilight my thoughts turned to Edwin Hubble, who first showed that some of those faint smudges in the sky were “island universes”—galaxies like our own. The universe was a far bigger place than we had heretofore imagined. As Hubble observed about this quest to understand:

From our home on the Earth, we look out into the distances and strive to imagine the sort of world into which we are born. Today we have reached far out into Space. Our immediate neighborhood we know rather intimately. But with increasing distance our knowledge fades, and fades rapidly, until at the last dim horizon we search among ghostly errors of observations for landmarks that are scarcely more substantial. The

search will continue. The urge is older than history. It is not satisfied and it will not be suppressed.

Hubble’s comment was about the observational difficulties of distance estimation and the challenges associated with identifying standard candles. However, it could just as easily have been a meditation on computing, for we are driven by our own insatiable desires for better algorithms, more flexible and reliable software, new sensors and data analytics tools, and by ever larger and more faster computers.

Computing the Future

Why do we compute? I suspect it is for at least two, related reasons, neither relating to publication counts, tenure, wealth, or fame. The first is the ability to give life to the algorithmic instantiation of an idea, to see it dance and move across our displays and devices. We have all felt the exhilaration when the idea takes shape in code, then begins to execute, sometimes surprising us in its unexpected behavior and complexity. Computing’s analogue of *deus ex machina* brings psychic satisfaction.

The second reason is that computing is an intellectual amplifier, extending our nominal reach and abilities. I discussed the power of computing to enable and enhance exploration in a previous blog entry. It is why those of us in computational science continually seek better algorithms and faster computer systems. From terascale to petascale and the global race to exascale, it is a quest for greater fidelity, higher resolution, and finer time scales. The same deep yearning drives astronomers to seek higher resolution detectors and larger telescope apertures. We are all chasing searching the ghostly signals for landmarks.

It is our ability to apply our ideas and their embodiment in code to a dizzying array of problems—from the prosaic to the profound—that attracts and compels us. It is why we compute. Hubble was right. We compute because we want to know and understand. The urge is deep and unsatisfied. It cannot be denied.

Jason Hong is an associate professor of CS at Carnegie Mellon University. Daniel Reed is vice president of Technology Strategy & Policy and the eXtreme Computing Group at Microsoft.

© 2013 ACM 0001-0782/13/03

UIST 2013

26th ACM Symposium on
User Interface Software & Technology

October 8th - 11th, 2013
St Andrews, Scotland, UK

Paper Deadline: April 5

Co-located with ITS 2013: ACM Interactive Tabletops and Surfaces (Oct 6th - 9th)



<http://acm.org/uist>

Poster designed by rareloop.com

Decoding Dementia

Computer models may help neurologists unlock the secrets of brain disorders, from Alzheimer's to cancer.

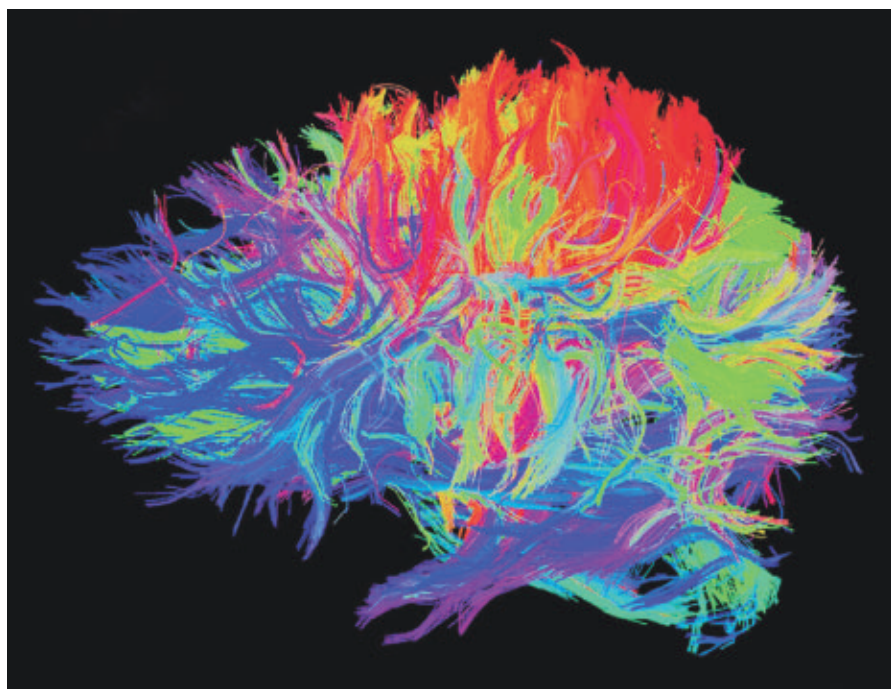
THERE IS, AS yet, no cure for Alzheimer's disease, and little in the way of treatment. But computer models that can predict the course of the illness, which gradually destroys memory and other cognitive functions, might allow doctors to manage the disease and perhaps help scientists to better understand it. Computer modeling might, in fact, lead to clearer prognoses and better treatment for a whole range of brain disorders, from Parkinson's disease to brain cancer.

One group of scientists, from Weill Cornell Medical College in New York and the University of California, San Francisco, developed a model that starts with a magnetic resonance image of the early stages of Alzheimer's or frontotemporal dementia and predicts how the illnesses would spread throughout the brain.

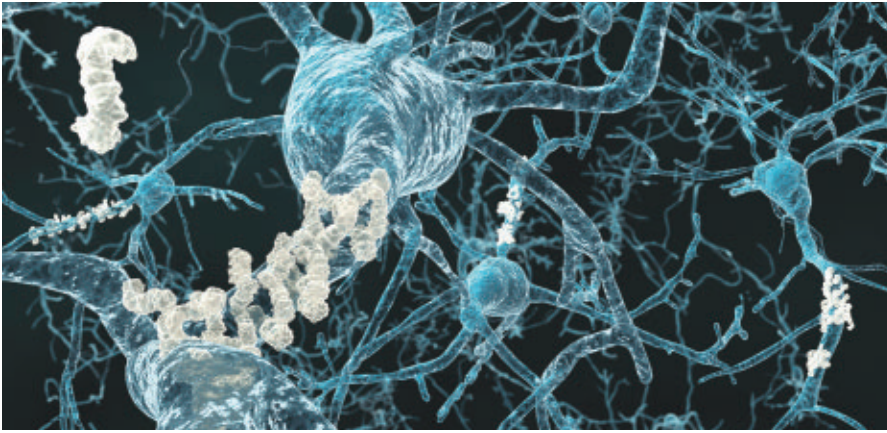
"We assume that the disease in the brain diffuses from one place to another as if it were a gas," says Ashish Raj, assistant professor of computer science in radiology at Weill Cornell. But instead of spreading through open air, the "gas" in question had to travel along the brain's neural pathways. "We had to come up with the right kind of math that would model how something diffuses in a network rather than in space."

The model works on the idea, arising from previous studies by other scientists, that these illnesses begin in one part of the brain and spread along a network of neurons. Scientists have recently learned that dementia seems to behave as if it is caused by an errant protein, known as a prion. Prions are known to cause the brain illness Creutzfeldt-Jakob disease, the human

form of mad cow disease. Proteins normally fold into shapes that dictate their function in the body, but prions are misfolded, and can transmit the misfolding to healthy proteins, causing them to form into plaques called amyloids that destroy brain tissue and seem to play a role not only in Alzheimer's but also in other degenerative brain diseases such as Huntington's and Parkinson's.



Brain fiber tracts (colors represents the orientation of the fibers) are used to study connectivity networks, whose diffusion dynamics model dementia.



The accumulation of amyloid plaques between neurons in the brain is one of the hallmarks of Alzheimer's disease.

Raj and Michael Weiner, professor of radiology and biomedical imaging at UCSF, mapped out how proteins are dispersed in healthy brains, then applied the assumption that faulty proteins would spread from one area to another. To determine the rate at which they would spread, they looked at how the prevalence of dementia varies by age, which should be related to how long the disease takes to develop. They ran the model, then compared the results to actual brain images from 18 patients with Alzheimer's and 18 with frontotemporal dementia, and found the model matched reality. Those results, gleaned by bootstrap analysis of their data, seems to validate the prion hypothesis, Raj says.

In practical use, the model should be able to look at a single scan from a dementia patient, figure out where he or she is in the course of their disease, and predict its progression. That might help doctors and patients make decisions about the few treatment options that exist.

Marc Diamond, associate professor of neurology at Washington University in St. Louis School of Medicine, says Raj's model agrees with his studies, in cells and mouse brains, of the molecular mechanisms for Alzheimer's spread. "The two are very consistent with one another," Diamond says. "They are basically saying the network involvement in some of these diseases is consistent with an agent that moves between neurons, based on how they're connected to each other."

The computer model has limitations, Diamond says; he calls it a "30,000-foot view of the whole pro-

cess" that does not explain why disease might not spread between some connected neurons. But he adds it might help provide researchers with hints as to regions of the brain they ought to study with other methods.

The MRI scans came from the Alzheimer's Disease Neuroimaging Initiative, run by Weiner, which is collecting brain images that may help scientists understand the course of the disease. The patients in question have only been followed for two to four years, and Raj says it will be helpful to have images that cover a decade or two. But even over that shorter time span, an incorrect model would diverge from the actual images, which this model did not, he explains.

One advantage of the model, he says, is its simplicity. It ignores whatever complex activities may be taking place at a cellular level and focuses on diffusion patterns to get an accurate picture of disease progress. That, says Raj, is an

The computer model might help provide researchers with hints as to regions of the brain they ought to study with other methods.

example of how computer scientists can sometimes aid other scientists.

Another advantage is that the model relies on well-understood principles. "We have used a set of tools from graph theory that really are quite well known, but have found a way that was quite unexpected to apply them to disease," Raj says. His group is not the only one using that sort of modeling to study the communications network in the brain and how it relates to dementia. Neurologists at VU University Medical Center in Amsterdam also applied graph theory to examine the differences in connectivity in the neural networks of healthy brains versus Alzheimer's brains.

They started with the idea that healthy brains may follow a small world network model, in which points in the network connect to several other points, and hubs with many connections link one cluster of connections to many others. What they found, says Willem de Haan, one of the researchers involved in the work, was that connections in the brains of Alzheimer's patients showed a less ordered series of connections than healthy brains. Then they looked at patterns of brain activity from electroencephalograms to see if there was any link between activity levels and areas that were damaged by amyloid plaques.

"The hub in the network, the well-connected points, seemed vulnerable for some reason," de Haan says. "We wanted to try to explain how these hub areas became vulnerable."

Operating on the hypothesis that areas of high neural activity were more likely to suffer the damage that leads to dementia, the group employed a neural mass model, a simplified model that focuses on only a few variables to describe the activity of a population, or mass, of neurons. They then applied an algorithm that "damaged" the masses, reducing the number of signal-transmitting synapses based on the amount of activity in that particular area. They then used graph theory to analyze the results, and found the outcomes matched those in Alzheimer's patients.

The computer model provides a theory to be tested in actual tissue, de Haan says. And it might suggest a possible treatment to ward off the early

effects of the disease. If increased brain activity does lead to structural damage, perhaps methods that alter activity levels, through the use of electrical or magnetic stimulation, might slow the disease's progression. The researchers would also like to look at scans of people in the early stages of Alzheimer's to see if they can find the differences in activity levels that the computer model predicts.

Another Path

Kristin Rae Swanson, a professor of neurological surgery at Northwestern University's Brain Tumor Institute, also combines brain scans and mathematical modeling to predict how areas of damage spread through the brain. But instead of misfolded proteins, she is looking at the spread of glioblastoma, a deadly brain tumor. All glioblastoma patients receive MRIs and are treated with radiation and perhaps brain surgery, but Swanson says her models can identify parts of the tumor that the scans cannot pick up, and predict how a tumor will spread through a particular patient's brain. That in turn could lead to individualized treatments that could prolong a patient's life and minimize side effects of excessive radiation treatment. "It's a different approach to personalized medicine," she says.

MRIs only pick up the densest part of a tumor and miss cells that are more

MRIs only pick up the densest part of a tumor and miss cells that are more sparsely concentrated.

sparsely concentrated; Swanson compares it to looking at an iceberg from above. The model takes advantage of what is known about the cancer's proliferation rate (how quickly the cells reproduce) and its diffusion rate (how quickly they move through the brain), which is faster in the white matter than the gray matter. It also takes into account that cells tend to move along the blood vessels and fiber tracts in the white matter. So by looking at the shape of the iceberg, her model can predict where the invisible parts lie, so that a doctor can target them. And by seeing which parts of the brain's structure are affected, it predicts how the tumor will grow.

Raj believes his model could also be used to study the dynamics of brain development, looking not at just

where brain connections go wrong, but how they form in the first place. It might also be applied to epilepsy, in which what is diffusing through the brain is not a misfolded protein but a disordered electrical signal. In fact, it should work with any brain ailment that is characterized by something spreading throughout the brain. **Q**

Further Reading

De Haan, W., Mott, K., Van Straaten, E.C.W., Scheltens, P., and Stam, C.J. Activity dependent degeneration explains hub vulnerability in Alzheimer's disease, *PLOS Computational Biology*, Aug. 16, 2012.

Duch, W. Computational models of dementia and neurological problems, *Methods Mol Biol.* 401, 2007.

Harpold, H.L.P., Alvord, E.C. Jr. and Swanson, K.R.

The evolution of mathematical modeling of glioma proliferation and invasion, *J Neuropathol Exp Neurol* 66, Jan. 2007.

Raj, A., Kuceyeski, A. and Weiner M.W. A network diffusion model of disease progression in dementia, *Neuron* 73, Mar. 22, 2012.

Weiner, M. Dr. Michael Weiner, MRI specialist, San Francisco Vets Admin <http://www.youtube.com/watch?v=aNo-KwmReiA>

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

© 2013 ACM 0001-0782/13/03

In Memoriam

Jim Horning, 1942–2013



James J. "Jim" Horning, an ACM Fellow recognized for his work in programming language design and specification

methodology, passed away on January 18, 2013 in Palo Alto, CA. He was 70 years old.

ACM President Vinton Cerf called Horning a "quintessential member" of the computer science community as well as the ACM community. A member since 1965, Horning's devotion to the organization was never

more evident than with his work over the last decade as co-chair—with Calvin C. (Kelly) Gotlieb—of the ACM Awards Committee. Their collaboration was instrumental in bringing global recognition to the ACM Awards program as a true measure of professional excellence and respect.

Horning was a founding member and chair of the University of Toronto's Computer Systems Research Group in 1969, a Research Fellow at Xerox's PARC, and a founding member at DEC's Systems Research Center. His security expertise was in great demand at companies such as

McAfee, SPARTA, InterTrust Technologies, and Silicon Graphics.

As the information age escalated, Horning became increasingly concerned about ensuring privacy, security, and trustworthiness in the private and public sector. He was an active member of ACM's Committee on Computers and Public Policy (CCPP) and ACM's Public Policy Council (USACM) from their inception. Peter Neumann, CCPP chair and moderator of the *ACM Forum on Risks to the Public in Computers and Related Systems*, called Horning "one my favorite friends, colleagues, associates,

and long-time inspiration." Their friendship spanned 38 years, with Horning contributing to the very first issue of the *Risks Forum* in August 1985. "He made many thoughtful technical and socially aware contributions, always with wisdom, common sense, and humanity."

Cerf recalled Horning as "the best one can ever find in our profession. Always ready to help with a quiet style of leadership, many of us felt free to consult him and frequently did for advice borne of experience and calm thought. I will miss him as will so many in our field and community."

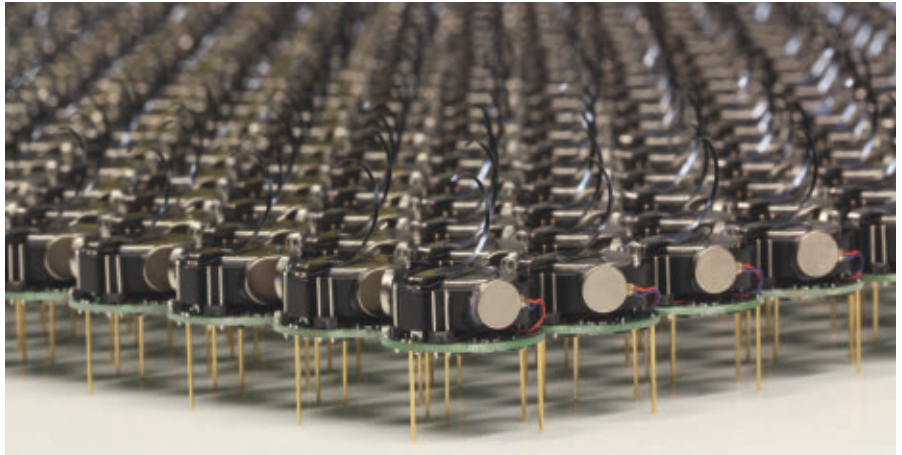
Rise of the Swarm

Guided by collective intelligence, teams of small, simple robots could soon accomplish amazing feats.

THE WATERS OFF the west coast of Scotland are lined with once rich, complex coral reefs. Over the years, bottom-trawling fishermen have all but ruined them, leaving the coral broken and displaced. Divers have begun working to repair the damaged reefs, but in some places they lie more than 650 feet below the surface, far too deep for a standard scuba diver. So, in August, a group of researchers at Heriot-Watt University in Edinburgh announced plans to build a fleet of underwater robots to do the job. These so-called corallbots will be able to work at greater depths than human divers and stay down for longer periods. The corallbots will work cooperatively, identifying dead chunks of the reef via high-resolution cameras, using the lifeless fragments to rebuild the inner structure, then stacking still-living corals on the outside.

Despite the complexity of their task, the robots themselves will not be all that bright, according to Heriot-Watt computer scientist David Corne. “You could rebuild the reefs without particularly complex planning or intelligence in the robots,” he says. Corne points to ants, termites, and wasps as examples. Each of the individual creatures follows a series of simple rules as they react to patterns in their environment, but as a group they can produce fantastic structures. “A termite mound has chambers and a whole ventilation network,” he notes. And yet it is built by a collection of simple intelligences.

Although Corne points to termites and wasps as proof the concept could work, he could just as easily cite the advances in numerous robotics labs across the world. Researchers have been interested in the notion of robots guided by swarm-type intelligence for decades, but in the last few years these collectives have become reality. Scientists have demonstrated more than a hundred robots working together at a time, and within the next year they hope



Harvard's Kilobot project was designed as a low-cost, scalable robot system for demonstrating collective behaviors.

to push the population of a controlled swarm past 1,000. “It’s unreal what we can do as a community compared to just five years ago,” says Magnus Egerstedt, a roboticist at the Georgia Institute of Technology in Atlanta. “There had been lots of simulated swarm robotics, but actually seeing 50 robots reliably doing things together now is not out of the question, and five years ago it was.”

Follow No Leader

Several years ago, while in Budapest, Hungary, for a conference, Egerstedt met a shepherd, and fell into a conversation about the role of the herding dog in directing a flock. Egerstedt himself had long been puzzled by a theoretically related question. “If you’re surrounded by a million robot mosquitoes, and you have a joystick, what do you actually do with the joystick?” he asks. “How should you interface with the swarm?” To him, the interaction of the shepherd with the herding dog seemed like the answer. “It seemed like a really natural way of thinking about human-swarm interactions.”

Back in Atlanta, Egerstedt planned an experiment to test the idea. He recruited test subjects to see how they would fare in directing a group of 25 simple, wheeled robots around the floor

of his lab. Each participant was given a joystick and asked to accomplish a number of relatively simple tasks, such as manipulating one of the robots to arrange the others in a circle or a wedge shape. The robots were programmed to react to one another, so the follow-the-leader technique seemed like it would work. Yet the participants failed miserably. “People were overall quite pathetic at it,” Egerstedt recalls.

As a result, Egerstedt moved away from this leader-based interaction to a more democratic approach. “If you’re surrounded by a million mosquitoes, you’re probably not going to pick a key mosquito and start dragging it around,” he acknowledges now. “You’d probably wave around in the air and try to get them to move around in some pattern.”

Egerstedt is now designing a follow-up experiment in which the robots move around within a network of wireless routers. This time, participants will be given a motion-capture wand and asked to accomplish similar tasks with the robots. The difference is that the wand will create flows instead of directing a particular robot. When a user waves the wand in one area of the network, the nearby routers will track that motion, and, as robots pass nearby, the routers will tell them to move the same

way. “Then the routers will also talk to each other in a distributed way and figure out how to conduct traffic,” he says.

Dispensable Machines

Although Egerstedt is excited about the potential of what he calls swarm conducting, he is also quick to point out the amazing work going on in other labs around the world. One oft-cited example of the possibilities of swarm-controlled machines is the work of roboticist Vijay Kumar and his team at the University of Pennsylvania. Kumar’s group has developed small quadrotors that fly in sync like flocks of birds. The robots vary pitch, roll, and yaw by adjusting the speed of their rotors, and roughly 100 times per second they calculate their position relative to one another and communicate their coordinates via radio.

Currently, the robots rely on a motion capture system in the lab that creates and updates a detailed map of the environment for each quadrotor. This allows the robots to remain small and light, since they do not have to carry bulky onboard sensors to do the mapping work themselves. But Kumar’s team has also demonstrated a larger flying robot outfitted with sensors that autonomously creates these virtual representations of the surrounding space. He says he can envision these robots exploring areas off-limits to humans. “Our interests are in search and rescue,” he says. “I envision a day when robots are the first ones on the scene.”

The robots will be able to function in dangerous environments in part because they are designed to be expendable. “The role of individual robots will be minimized,” Kumar explains. “The idea of one superior unit to control everybody is not necessary.” In fact, he says, you have to assume that a percentage of the robots will be lost in dangerous environments. The algorithms that control them need to be built with this in mind. “You need to make sure that the algorithms will work independent of the number of units.”

Expanding the Swarm

The need to push those algorithms is part of what drove Harvard University roboticist Mike Rubenstein to construct the Kilobot, an inexpensive robot no wider than a quarter that moves around

“Our interests are in search and rescue,” Kumar says. “I envision a day when robots are the first ones on the scene.”

on three simple, vibrating stick legs. Rubenstein’s work is linked to Harvard’s Robobees project—an NSF-funded multiyear effort to build a swarm of autonomous robotic bees that may eventually be able to pollinate areas like living bees do.

The Kilobots offer a way to test the same algorithms that will guide the bees, but in a simpler, cheaper, grounded machine. If the goal is to develop control algorithms that can handle true swarms, Rubenstein reasoned, then those algorithms need to be tested at scale. Software that effectively guides the actions of a few dozen robots could break down when dealing with a few thousand. So Rubenstein designed the cheapest mobile robot he could, a squat, cylindrical machine with an infrared sensor and transmitter and, in all, less than \$15 in parts. For the last few months he has been assembling his swarm with a goal of reaching 1,000 early this year.

In the meantime, Rubenstein and his collaborators have also demonstrated many insect-like behaviors such as navigation and foraging collective transport on swarms of up to 100 robots. Like Kumar, he views the individual as dispensable. “Everything you need to do can be done on the group level, not on the individual robot level,” he says. “You give them a program and they run it. User interaction is not necessary.” Still, if a human observer wanted to switch the swarm’s task midstream, it would be possible to communicate that change via a central controller. This central controller would not actively mediate the actions of the robots. It would merely set them to work, or pause their actions and send them new directives when necessary.

The insights gleaned from develop-

ing control systems for robots could also be transferred to other fields, according to Alcherio Martinoli, a roboticist at the Swiss Federal Institute of Technology in Lausanne. “You can use them as a testbed for fine-tuning certain methods and then you can transport the methods,” he says. For example, Martinoli and his colleagues are exploring how their algorithms might assist a flight traffic control system should our future skies become overcrowded with piloted and unmanned planes.

Still, it is the direct applications, and the tiny machines themselves, that draw so much attention to the field of swarm robotics. Martinoli has explored using swarms to inspect complex industrial equipment such as jet turbines. Kumar’s robots could scour dangerous disaster zones or crumpled buildings for survivors. Harvard’s Justin Werfel, a colleague of Rubenstein, is leading a project, TERMES, that aims to create a fleet of cooperative construction robots. Werfel envisions these machines building bases in extreme locales such as the deep sea or the Moon. With each application, numerous technological hurdles remain, but experts are quick to note that the basic idea of a swarm of relatively unintelligent, independent agents working together to achieve a complex goal is not a stretch at all. “We know it can be done because we see it happening in nature,” Werfel says. ■

Further Reading

Bonabeau, E.

Swarm Intelligence: From Natural to Artificial Systems, Oxford Univ., 1999.

Kushleyev, A., Mellinger, D. and Kumar, V. Towards a swarm of nanoquadrotors; <http://www.youtube.com/watch?v=YQIMGV5vtd4>

Martínez, S., Cortés, J. and Bullo, F. Motion coordination with distributed information, *IEEE Control Systems Magazine*, Aug. 2007.

Olfati-Saber, R., Fax, J.A. and Murray, R.M. Consensus and cooperation in networked multi-agent systems, *Proceedings of the IEEE*, Jan. 2007.

Rubenstein, M., Ahler, C. and Nagpal, R. Kilobot: A low cost scalable robot system for collective behaviors, *IEEE Intl. Conf. on Robotics and Automation*, 2012.

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

© 2013 ACM 0001-0782/13/03

Cybercrime: It's Serious, But Exactly How Serious?

Symantec says \$110 billion annually while McAfee says \$1 trillion. Why can't anyone agree?

EVERYONE AGREES CYBERCRIME affects everyone—governments, corporations, the public—but to what extent? And while vast sums are spent on security to protect against the evildoers, why is it so difficult to determine the amount of the damage they have done?

According to its most recent study, security software manufacturer Symantec Corp. reports cybercrime is costing the world \$110 billion every year. But, according to McAfee Inc.—Symantec's closest competitor—the actual annual cost worldwide is almost 10 times that, approximately \$1 trillion.

What's going on here?

Unfortunately, say security experts, there seem to be at least four hurdles to accurate reporting:

- **Failure to report.** Many organizations that have been cybercrime victims do not want to report the problem because they perceive it as bad for business.

- **Self-selection bias.** Organizations that have not detected losses may be more likely to respond to cybercrime surveys than those that have. Or those that have had very public large losses may be more prone to reply than those with moderate unreported losses.

- **No standard mechanism for accounting for losses.** Sometimes downtime is figured into the mix. Sometimes the cost of buying new equipment or upgrades or security services or outside consultants is included in the total. There is no agreement on what and what not to include.

- **Undetected losses.** Often organizations are not even aware they have had losses or the full magnitude of the crime is not known.

Consider, for example, a company doing advanced research has a break-in and proprietary information is copied out of its computers, says Eugene H. Spafford, a professor of computer



science at Purdue University. "If the company discovers the intrusion—and it might not—an audit might determine the loss equals the cost cleanup and perhaps changing to new security software. But what if the company isn't aware of all that was taken or doesn't know how to evaluate it? And what if that same proprietary information shows up a year later in a competitor's product in another country? How then do you evaluate the loss? And what if the product has national defense associated with it? How do you put a value on a significant enough product? Millions of dollars? Billions?"

McAfee attributes a portion of the discrepancy between its reporting and Symantec's to the fact that its study focuses on the amount of money businesses lose worldwide due to both malicious and accidental data loss.

The company concedes, however, that coming up with an estimate is particularly difficult because there are so

many facets that need to be considered.

"You need to add up losses due to corporate espionage, losses that can't be quantified, losses from damage to a brand's reputation, and so on," says a McAfee spokesperson. "But the hardest to estimate is the cost to the long-term competitiveness of the U.S. economy. What is the cost to the U.S. down the road when competitors to our best hardware, software, and bio-tech companies emerge in the future and take away market share and American jobs? Those costs—which could be huge—are the ones that are the most difficult to evaluate."

And, for consumers, it is not only about what is lost through fraud with online banking; it is also about their digital assets.

"What's the worth of that book draft they've been working on for a year?" asks the McAfee spokesperson. "And how much are their photographs worth? What about the worth of their online

identity? Most victims spend a considerable amount of time trying to recover their identities and recreate information they've lost. What is that time worth?"

The true cost of cybercrime, he adds, involves looking at all these questions and adding them up "using a strong, clear, defensible methodology." Many companies and think tanks do not have the time or the money to do that kind of extensive research, he says.

Symantec chose not to comment or participate in this story.

Meanwhile, Cormac Herley, principal researcher at Microsoft Research, says he has "no faith whatsoever that either one of the numbers—Symantec's or McAfee's—is anywhere close to the truth. You can call anything an estimate," he says, "but that doesn't mean it's a reasonable reflection of the underlying reality."

Herley and his co-researcher, Dinei Florencio, recently wrote a paper, "Sex, Lies and Cybercrime Surveys," after reading cybercrime estimates "that varied by orders of magnitude. I mean, many things have some wiggle room. But if physicists couldn't agree on the speed of light to within four orders of magnitude, they would just confess they didn't know."

Herley blames the methodologies in the cybercrime surveys that, he says, almost always exaggerate the numbers on the high side. He believes the actual numbers are far smaller.

The problem, he says, is that cybercrime surveys are not like voting surveys where everyone's answer counts equally.

"When you ask people what they lost from cybercrime, you have no ability to verify that they understood the question and that they answered truthfully," he explains. "And then, when even a single person gives you a number that is grossly incorrect, they have the ability to destroy the entire survey. It almost always results in a major upward bias in the numbers."

To illustrate how one person can make nonsense out of a survey, Herley suggests a study to determine how many people have pet unicorns. "If you ask 100 people (which substitutes for a population of 100 million people in the country), it means that whatever number you get you need to multiply by one million. Then you conduct

Cynics have charged that cybercrime stats are artificially inflated to scare more people into buying security software.

the survey and everyone truthfully answers "zero," except for one person who misunderstands the question and says that, yes, they have one unicorn because their daughter has a stuffed one in her bedroom. Your estimate now shows there are one million unicorns in the U.S. It's completely incorrect and it's based on that one incorrect answer."

If that is the case, does it even make sense to try and determine the cost of cybercrime given the likelihood the results will be hugely inflated? Experts say "yes;" that if an organization uses the same consistent method repeatedly, trends emerge and that is valuable for those battling cyber losses.

In addition, from an awareness standpoint, experts say it is important to get the business world, private individuals, and government organizations to understand the magnitude of the problem. Otherwise, the usual attitude is "we've never had a problem so it's likely we won't have one in the future."

Cynics have charged that cybercrime stats are artificially inflated to scare more people into buying security software. And, they suggest, companies that profit by selling anti-malware software should not be the ones reporting on the size of the malware problem.

On the other hand, say observers, who else is going to conduct analyses of security other than the security companies who know the field, know whom to ask, and generally have respected names so people are likely to respond to them with good information.

"You're not likely to see a survey in this area conducted by Hostess Snack Foods," said one. "As for the government doing it, many organizations simply don't want to report to the government that they've had losses be-

ACM Member News

HANAN SAMET, A TRAILBLAZER IN SPATIAL DATABASES



When Apple CEO Tim Cook found it necessary to apologize for the quality of Apple Maps and

iPhone users began using Google Maps instead, it underscored the importance of the pioneering work Hanan Samet has been doing on spatial information for the past 36 years. In fact, his recent paper, "Duking It Out at the Smartphone Mobile App Mapping API Corral: Apple, Google, and the Competition," won a "best paper" award at the recent 1st ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems.

Samet, a professor of computer science at the University of Maryland, says he is particularly honored to have won the 2011 Paris Kanellakis Theory and Practice Award for his research on quadrees and other multidimensional spatial data structures for sorting spatial information. "It is well known that leading vendors such as Google and Microsoft use Hanan's results in their GIS and commercial mapping systems," notes Dinesh Manocha, a CS professor at the University of North Carolina. "He can be regarded as the world's leading authority on spatial databases and multidimensional data structures." Samet referred to Kanellakis as "a friend who reached out to me when he heard of my work and involved me in the 1992 PODS conference after I co-authored the first paper on spatial data mining in the 1990 PODS Conference." He quipped that he was proud to have seemingly solved problems that were deemed unsolvable "primarily because I did not know they could not be solved."

Samet is currently working on building spatial indices based on textual specifications of spatial data, in contrast to geometric ones, for enabling text and tweets to be accessed with a map query interface.

—Paul Hyman

cause they don't trust how that information will be used."

But Microsoft's Herley says he believes "with very high confidence and without much fear of contradiction the methodologies the companies use produce bogus answers. As far as their motivations go, I just don't know and I don't want to speculate. Mistakes happen all the time even when there's no intent to deceive."

However, Ross Anderson suspects that most cybercrime statistics—"like the ridiculous \$1 trillion number which means cybercrime is 2% of the world's GDP"—have been unreliable "because people compiling them (like policemen or security software vendors) have had some axe to grind." Anderson is professor of security engineering at the University of Cambridge's Computer Laboratory.

His 2008 study, "Security Economics And The Single Market," reports that has been the case for years. And in his more recent paper, "Measuring The Cost of Cybercrime," he suggests society ought to spend less money on anti-virus software and more on policing the Internet.

"Many cybercrimes are committed by a small number of people," he says. "For example, in 2010, a third of all the spam in the world was sent by one botnet. So it would be a lot more efficient to just arrest the bad guys and put them in jail than to expect several hundred million users worldwide to run anti-virus and anti-spam software. Of course the anti-virus and anti-spam companies don't agree."

A large part of the true cost of cyber-

"Many organizations simply don't want to report to the government that they have had losses because they don't trust how that information will be used."

crime is the money the world spends on anti-virus software, he maintains, adding "in fact, the anti-virus companies make much more money out of spam than the bad guys do."

If, in fact, measuring the true cost of cybercrime is viewed as important, experts have recommendations.

Purdue's Spafford suggests that a reasonable set of metrics—and a reasonable set of questions to obtain those metrics—needs to be devised by an organization familiar with creating surveys and calculating costs.

He recommends a coalition of software or hardware vendors, perhaps one that already exists, perhaps an organization like the National Institute of Standards and Technology (NIST).

"Whoever it is," he says, "it needs to be someone who has everyone's trust. And that's not going to be easy, nor is it going to be cheap."

Cambridge's Anderson suggests

taking a different route: "Stop wasting money on measuring cybercrime and stop wasting money on cyberwar," he says. "Spend it on the police instead." ■

Further Reading

Anderson, R.

Measuring the cost of cybercrime, <http://www.cam.ac.uk/research/news/how-much-does-cybercrime-cost/>, June 8, 2012.

Flores, D. and Herley, C.

Sex, lies and cyber-crime surveys, Workshop on Economics of Information Security, <http://research.microsoft.com/apps/pubs/default.aspx?id=149886>, June 2011.

HouseResource.org

Cybersecurity: Assessing the Immediate Threat to the United States (video), <http://www.youtube.com/watch?v=Tmm-rv6oTLY>, posted May 27, 2011.

McAfee

Unsecured economies: Protecting vital information, http://www.cerias.purdue.edu/assets/pdf/mfe_unsec_econ_pr_rpt_fnl_online_012109.pdf, Jan. 21, 2009.

Symantec

2012 Norton Cybercrime Report, http://now-static.norton.com/now/en/pu/images/Promotions/2012/cybercrimeReport/2012_Norton_Cybercrime_Report_Master_FINAL_050912.pdf, May 9, 2012.

Verizon RISK Team

2012 Data Breach Investigations Report, http://www.verizonbusiness.com/resources/reports/rp_data-breach-investigations-report-2012-ebk_en_xg.pdf, 2012.

WhiteHouse.gov

New Initiatives To Combat Cyber Terrorism (video), <http://www.youtube.com/watch?v=FyRMfZXPxLA>, posted May 30, 2012.

Paul Hyman is a science and technology writer based in Great Neck, NY.

© 2013 ACM 0001-0782/13/03

Milestones

Computer Science Awards, Appointments

CERF NSB APPOINTEE

Last January, President Obama announced his intention to appoint Vinton G. Cerf to the National Science Board. Cerf, Vice President and Chief Internet Evangelist at Google and ACM President, is an appointee to the 25-member National Science Board, which is the governance body for the National Science Foundation, and additionally serves as an independent body

of advisors to both the President and the U.S. Congress on policy matters related to science and engineering and education in science and engineering.

FCC CHAIRMAN'S AWARD

The Federal Communications Commission (FCC) honored Juan Gilbert and a team of students from Clemson University's Human-Centered Computing division with the FCC Chairman's

2012 Award for Advancement in Accessibility. The team was recognized for their Prime III: A Universally Designed Voting Machine that enables voters with disabilities to cast votes in a private, secure environment without assistance.

BBVA AWARD TO ZADEH

Lotfi A. Zadeh received the BBVA Foundation Frontiers of Knowledge Award in the

Information and Communication Technologies (ICT) category.

Zadeh was recognized for the invention and development of fuzzy logic, a breakthrough cited as enabling machines to work with imprecise concepts, in the same way humans do, and thus secure more efficient results more aligned with reality. In the last 50 years, this methodology has generated over 50,000 patents in Japan and the U.S. alone.

ACM Fellows Inducted

ACM HAS RECOGNIZED 52 of its members for their contributions to computing that are fundamentally advancing technology in healthcare, cybersecurity, science, communications, entertainment, business, and education. The 2012 ACM Fellows personify the highest achievements in computing research and development from the world's leading universities, corporations, and research labs, with innovations that are driving economic growth in the digital environment.

"These men and women are advancing the art and science of computing with enormous impacts for how we live and work," said ACM President Vinton G. Cerf. "The impact of their contributions highlights the role of computing in creating advances that range from commonplace applications to extraordinary breakthroughs, and from the theoretical to the practical. Some recipients have also helped to broaden participation in computing, particularly among underrepresented groups, and to expand its impact across multiple disciplines."

The ACM Fellows Program was established by Council in 1993 to recognize and honor outstanding ACM members for their achievements in computer science and information technology and for their significant contributions to the mission of the ACM. For a complete list of ACM Fellows, visit <http://fellows.acm.org/>

2012 ACM Fellows

Gustavo Alonso

ETH Zurich (Swiss Federal Institute of Technology)

Lars Arge

Aarhus University

Pierre Baldi

University of California, Irvine

Hans-J. Boehm

Hewlett-Packard

Craig Boutilier

University of Toronto

Tracy K. Camp

Colorado School of Mines

Rick Cattell

Cattell.Net LLC

Larry S. Davis

University of Maryland

Ahmed K. Elmagarmid

Qatar Computing Research Institute

Wenfei Fan

University of Edinburgh

Lixin Gao

University of Massachusetts, Amherst

Simson Garfinkel

Naval Postgraduate School

Garth A. Gibson

Carnegie Mellon University

Saul Greenberg

University of Calgary

Markus Gross

ETH Zurich (Swiss Federal Institute of Technology)

David P. Grove

IBM Research

Jonathan Grudin

Microsoft Research

Rachid Guerraoui

EPFL (École Polytechnique Fédérale de Lausanne)

Manish Gupta

Goldman Sachs

John Hersberger

Mentor Graphics Corporation

Andrew B. Kahng

University of California, San Diego

Anna Karlin

University of Washington

Srinivasan Keshav

University of Waterloo

Gregor Kiczales

The University of British Columbia

Masaru Kitsuregawa

The University of Tokyo

Leonid Libkin

University of Edinburgh

Tova Milo

Tel Aviv University

Klara Nahrstedt

University of Illinois at Urbana-Champaign

Joseph O'Rourke

Smith College

Benjamin C. Pierce

University of Pennsylvania

Keshav K. Pingali

University of Texas, Austin

Andrew M. Pitts

University of Cambridge

Rajeev R. Rastogi

Amazon

Raj Reddy

Carnegie Mellon University

Keith Ross

Polytechnic Institute of NYU

Karem Sakallah

University of Michigan

Robert S. Schreiber

Hewlett-Packard

Steven Scott

NVIDIA

Bart Selman

Cornell University

Ron Shamir

Tel Aviv University

Yoav Shoham

Stanford University

Joseph Sifakis

EPFL (École Polytechnique Fédérale de Lausanne)

Alistair Sinclair

University of California, Berkeley

Clifford Stein

Columbia University

Ion Stoica

University of California, Berkeley

Roberto Tamassia

Brown University

Walter F. Tichy

Karlsruhe Institute of Technology

Patrick Valduriez

INRIA and LIRMM

Leslie Valiant

Harvard University

Kathy Yelick

University of California at Berkeley/Lawrence Berkeley National Laboratory

Ramin Zabih

Cornell University

Xiaodong Zhang

The Ohio State University



Association for Computing Machinery

Global Reach for Global Opportunities in Computing



Dear Colleague,

Today's computing professionals are at the forefront of the technologies that drive innovation across diverse disciplines and international boundaries with increasing speed. In this environment, ACM offers advantages to computing researchers, practitioners, educators and students who are committed to self-improvement and success in their chosen fields.

ACM members benefit from a broad spectrum of state-of-the-art resources. From Special Interest Group conferences to world-class publications and peer-reviewed journals, from online lifelong learning resources to mentoring opportunities, from recognition programs to leadership opportunities, ACM helps computing professionals stay connected with academic research, emerging trends, and the technology trailblazers who are leading the way. These benefits include:

Timely access to relevant information

- *Communications of the ACM* magazine
- *ACM Queue* website for practitioners
- Option to subscribe to the *ACM Digital Library*
- ACM's **50+ journals and magazines** at member-only rates
- *TechNews*, tri-weekly email digest
- *ACM SIG conference* proceedings and discounts

Resources to enhance your career

- **ACM Tech Packs**, exclusive annotated reading lists compiled by experts
- **Learning Center** books, courses, podcasts and resources for lifelong learning
- Option to join **37 Special Interest Groups (SIGs)** and **hundreds of local chapters**
- **ACM Career & Job Center** for career-enhancing benefits
- *CareerNews*, email digest
- **Recognition of achievement** through Fellows and Distinguished Member Programs

As an ACM member, you gain access to ACM's worldwide network of more than 100,000 members from nearly 200 countries. ACM's global reach includes councils in Europe, India, and China to expand high-quality member activities and initiatives. By participating in ACM's multi-faceted global resources, you have the opportunity to develop friendships and relationships with colleagues and mentors that can advance your knowledge and skills in unforeseen ways.

ACM welcomes computing professionals and students from all backgrounds, interests, and pursuits. Please take a moment to consider the value of an ACM membership for your career and for your future in the dynamic computing profession.

Sincerely,

Vint Cerf

President

Association for Computing Machinery



Association for
Computing Machinery

Advancing Computing as a Science & Profession



Association for
Computing Machinery

Advancing Computing as a Science & Profession

membership application & digital library order form

Priority Code: AD13

You can join ACM in several easy ways:

Online

<http://www.acm.org/join>

Phone

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

Fax

+1-212-944-1318

Or, complete this application and return with payment via postal mail

Special rates for residents of developing countries:

<http://www.acm.org/membership/L2-3/>

Special rates for members of sister societies:

<http://www.acm.org/membership/dues.html>

Please print clearly

Name _____

Address _____

City _____

State/Province _____

Postal code/Zip _____

Country _____

E-mail address _____

Area code & Daytime phone _____

Fax _____

Member number, if applicable _____

Purposes of ACM

ACM is dedicated to:

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

I agree with the Purposes of ACM:

Signature _____

ACM Code of Ethics:

<http://www.acm.org/about/code-of-ethics>

choose one membership option:

PROFESSIONAL MEMBERSHIP:

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

STUDENT MEMBERSHIP:

- ☐ ACM Student Membership: \$19 USD
- ☐ ACM Student Membership plus the ACM Digital Library: \$42 USD
- ☐ ACM Student Membership PLUS Print CACM Magazine: \$42 USD
- ☐ ACM Student Membership w/Digital Library PLUS Print CACM Magazine: \$62 USD

All new ACM members will receive an
ACM membership card.

For more information, please visit us at www.acm.org

Professional membership dues include \$40 toward a subscription to *Communications of the ACM*. Student membership dues include \$15 toward a subscription to *XRDS*. Member dues, subscriptions, and optional contributions are tax-deductible under certain circumstances. Please consult with your tax advisor.

RETURN COMPLETED APPLICATION TO:

Association for Computing Machinery, Inc.
General Post Office
P.O. Box 30777
New York, NY 10087-0777

Questions? E-mail us at acmhelp@acm.org
Or call +1-800-342-6626 to speak to a live representative

Satisfaction Guaranteed!

payment:

Payment must accompany application. If paying by check or money order, make payable to ACM, Inc. in US dollars or foreign currency at current exchange rate.

☐ Visa/MasterCard ☐ American Express ☐ Check/money order

☐ Professional Member Dues (\$99 or \$198) \$ _____

☐ ACM Digital Library (\$99) \$ _____

☐ Student Member Dues (\$19, \$42, or \$62) \$ _____

Total Amount Due \$ _____

Card # _____

Expiration date _____

Signature _____



DOI:10.1145/2428556.2428564

Pamela Samuelson

Legally Speaking

A Copyright Challenge to Resales of Digital Music

A currently pending case will have significant implications for secondary markets in digital goods.

HAVE YOU EVER purchased music from iTunes? If you no longer listen to certain songs or no longer like the band, you might want to resell those tunes. But is it lawful to do so? Capitol Records says no in a lawsuit it brought against ReDigi, Inc., whose platform enables resales of digital music to take place. To technologists, iTunes music might seem like an environmentally friendly substitute for CDs, but the law may see things differently.

Both Capitol and ReDigi filed motions for summary judgment in July 2012 to resolve their dispute. (A judge can grant summary judgment when there are no facts in dispute requiring a trial to determine who is right and when the only issue is how the law should apply to the undisputed facts.) The presiding judge heard oral argument on these motions in October 2012 and is likely to rule on them soon.

Because the case presents some novel legal issues, it is difficult to predict the outcome. Whatever the trial judge rules, though, this case will un-

doubtedly go to the Second Circuit Court of Appeals, and maybe even to the U.S. Supreme Court.

The stakes could hardly be higher for all of us who have purchased digital copies of copyrighted works. If Capitol wins, secondary markets for digital goods will be illegal. And resales of digital music or e-books, even among friends, would be as illegal as peer-to-peer file sharing of copyrighted content (unless that content is covered by a Creative Commons license).

ReDigi's Service

ReDigi provides a platform through which owners of digital music purchased from iTunes can resell that music to other music lovers. To initiate this process, prospective resellers must download ReDigi software and designate files they want to resell. ReDigi's software checks to make sure the files are eligible for transfer (by verifying the files were purchased from iTunes) and then migrates the data files for that music from the reseller's computer to cloud storage. ReDigi's client-side software deletes the data from the resell-

er's hard drive as the data is migrated to cloud storage.

Once in the cloud, that music is available for purchase by other ReDigi users. Those who purchase the music can maintain it in personal lockers in the cloud and access it there through streaming. ReDigi makes no new copy of the resold music; it simply updates its database about who owns that music. Alternatively, a purchaser can download resold music to his or her computer. Downloading music purchased through ReDigi migrates the digital file from cloud storage to the purchaser's hard drive.

ReDigi keeps a share of the resale price and provides credits to resellers so they can purchase more music from other resellers on the ReDigi platform. ReDigi also encourages its users to make more music available through the service by offering discounts on future purchases or prizes to those who use the service to buy music.

Capitol's Complaint

Capitol's main claim is that ReDigi makes infringing copies of sound re-



cordings in which Capitol owns copyrights: One copy is made in transmitting the music to the cloud, another when storing the music in the cloud, and a third when purchasers download the resold music.

A second claim is that ReDigi infringes Capitol's exclusive right to distribute copies to the public. Capitol argues that this occurs when ReDigi software transmits music files from the reseller's computer to the cloud and then when the ReDigi software transmits those files to purchasers.

In addition, Capitol asserts that ReDigi is indirectly liable for infringements committed by its customers who resell and purchase music through ReDigi's services. The uploaders are said to be unlawfully distributing the music, while those who download or access music in the cloud are alleged to be making unlawful copies of the music. ReDigi induces these user infringements, knowingly contributes to them, and financially benefits from infringements it could have prevented. Capitol charges that "ReDigi is [] a clearing-house for infringement and [has] a

business model built on widespread, unauthorized copying of sound recordings" owned by Capitol and others.

Capitol alleged infringement of copyrights it owns in music found on ReDigi's site. Tracks by Coldplay, Katy Perry, Lady Antebellum, Lily Allen, KT Tunstall, and Norah Jones were, for instance, found there.

ReDigi's First Sale Defense

It is common knowledge that when we buy a CD of recorded music, a book, or a DVD movie, we are free to resell that copy to anyone we choose for whatever

price we are able to get. We can also lend out the copy, give it as a gift, or throw it away if we get tired of it. That is because our personal property rights in the artifact we purchased override the copyright owner's rights to control further distribution of that artifact.

A more legalistic way to make this point is to say that copyright law grants owners the right to control distribution of copies to the public, but that right extends only to the first sale of that copy to the public. After that first sale, the copyright owner's right to control distribution of copies is said to be "exhausted." (Internationally, this principle is known as "exhaustion of rights" because any transfer of title to a copy of a copyrighted work—whether by sale, gift, or bequest—exhausts the copyright owner's distribution right.)

The exhaustion rule enables bookstores, libraries, video rental stores, and used CD stores to operate free from copyright owner control. Flea markets and Salvation Army stores also benefit from the exhaustion rule when they resell books, CDs, and DVDs.

Capitol alleged infringement of copyrights it owns in music found on ReDigi's site.

ReDigi believes the exhaustion rule applies to digital music purchased from iTunes as well as to CDs. It has designed its software and service to conform as closely as possible, given the unique characteristics of digital technologies, to the contours of the first sale rule. “Congress has not excluded digital files from resale,” ReDigi has argued. “Capitol has received the benefit that the limited monopoly of copyright protection provides when it was paid for the first sale.” To allow it to control resales of digital music would be an unwarranted extension of the copyright monopoly.

ReDigi relies on a 1973 case, *C.M. Paula v. Logan*, in which the defendant, after purchasing copies of Paula’s greeting cards, used a technical process to transfer the designs onto ceramic tiles for resale as artwork. Paula argued that this infringed its reproduction right, but the court ruled that no duplication had taken place. The copy Logan purchased was simply transferred to another medium.

Capitol’s Responses

Capitol’s rebuttals to ReDigi’s first sale defense mainly focus on technicalities of U.S. copyright law. As statutorily codified, the exhaustion of rights rule limits the distribution right, not the reproduction right. Insofar as ReDigi makes or encourages the making of unauthorized copies of digital music, Capitol asserts the exhaustion rule is inapplicable.

Capitol also claims that ReDigi is ineligible for an exhaustion defense because ReDigi does not own the music being resold. Even if the purchaser of music from iTunes owns that music, he or she is not reselling to ReDigi but rather to other ReDigi customers. ReDigi thus cannot rely on the exhaustion defense.

Although ReDigi claims to resell “used” music, Capitol says there is “no such thing as a ‘used’ digital file, akin to a dog-eared book or scratched CD because digital works can only be uploaded and transmitted in new copies...embodied on different disks or servers. Those copies are, in Capitol’s view, infringing because purchasers “get pristine copies of song files for less than they would pay through legitimate channels.”

Those who want to traffic in pirated files will not find a haven on ReDigi’s platform.

Capitol characterizes as “semantic machinations” and “technological smokescreens” ReDigi’s characterization of its process of transferring digital music files from one computer to another as migrating data by moving it in blocks to other computer memory. Migrating data involves copying, plain and simple, which Capitol claims to be entitled to control.

Finally, Capitol asserts that ReDigi cannot be sure that resellers of digital music have not saved the tunes on some other device. Capitol relies on a 2001 U.S. Copyright Office report that concluded that the exhaustion principle should not apply to transmission of digital copies because it poses an unreasonable risk of copying that would harm the market for copyrighted works.

So Who Is Right?

There is merit in ReDigi’s argument that it has gone “to great lengths to build a system that was compliant with copyright law in every respect.” Having determined that existing technologies were not suitable for such compliance, it developed technologies capable of operating within the constraints of the law. Its system does not allow duplication of copies, nor of dissemination of multiple copies, which ReDigi agrees would infringe copyright. ReDigi only facilitates the migration of data files from the reseller’s computer to the purchaser’s computer, which is consistent with the spirit, if not the strict contours, of copyright law.

Also meritorious is ReDigi’s claim that its “entire business model is to provide incentives for legally purchasing music.” It has built in numerous

checks on abuse of the system. Its terms of service forbid users from infringing copyrights; its software checks to ensure files designated for resale were lawfully acquired; it provides tools through which users can get rid of files they downloaded illegally; and it gives consumers the opportunity to acquire music lawfully. Moreover, ReDigi’s customers cannot cash out of the system; they must spend their credits within ReDigi on other music. Those who want to traffic in pirated files will not find a haven on ReDigi’s platform.

Whether Capitol or ReDigi wins on summary judgment will likely turn on whether the courts take a strict statutory approach to the exhaustion rule or construe the rule in light of its historical purposes.

Under the strict approach, ReDigi would seem ineligible for an exhaustion defense because copies made to facilitate resales do not fall within the current statutory language and ReDigi itself does not own copies whose resale it enables.

Under a purposeful approach, ReDigi has a far better chance of success because it has designed its system so there is “virtually no chance” that infringement would occur through use of its system “in anything more than isolated instances.” Moreover, the principle that one can resell property one has purchased runs deep in American law.

Capitol also does not seem to comprehend how computers work. Digital music residing on a hard drive is copied innumerable times as it is used, and it is often shifted around from one part of memory to another as the computer carries out different operations. This kind of copying is not what copyright law worries about.

ReDigi is surely right that this case will set a precedent that will have profound implications for the continued existence of secondary markets in digital goods and for the lawfulness of any resale or lending of copyrighted works in digital form. Stay tuned. This case really matters. 

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

Copyright held by author.

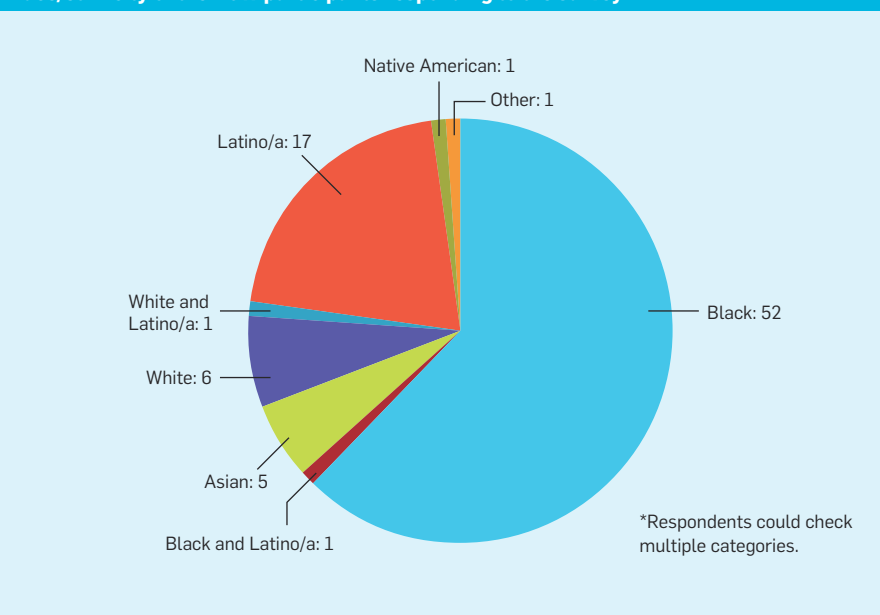
Broadening Participation Academic Careers Workshop for Underrepresented Groups

A longitudinal evaluation of the application of knowledge, skills, and attitudes of ACW participants.

A PRIMARY GOAL of Academic Careers Workshop (ACW) is to mentor minorities and persons with disabilities about the academic career ladder. This goal is accomplished through an annual workshop spanning several days that includes a series of panels and professional development sessions. These components culminate in a comprehensive experience designed to facilitate the professional trajectories of advanced doctoral students, early-career Ph.D.'s, and tenure-track faculty from assistant to associate and full professor/senior administrators. Alternative careers for doctoral-level computer scientists are encouraged and supported as well.

Beginning in 2005, the ACW was sponsored by Texas A&M University. In 2007, NSF began funding the ACW. In 2010, the scope was expanded to include persons with disabilities.¹ The 2012 Academic Careers Workshop (ACW) was organized by four groups: The Center for Minorities and People with Disabilities in Information Technology (CMD-IT); the Computing Alliance for Hispanic Serving Institutions (CAHSI); the Coalition to Diversify Computing (CDC); and the Alliance for Access to Computing Careers (Access-Computing). Among their goals, each organization is dedicated to increasing the number of people of color and

Race/ethnicity of the ACW participants responding to the survey.*



those with disabilities in graduate degree programs and academic careers in computing fields.

Not surprisingly, racial/ethnic minorities and persons with disabilities remain underrepresented in computer science and information technology disciplines particularly among the ranks of tenure-track faculty at colleges and universities. The Survey of Earned Doctorates (SED) is conducted by National Opinion Research Center (NORC).² According to the SED, 37 Black/African Americans and 48 His-

panics earned a doctorate degree in computer science in 2010. These numbers represent 2.2% and 2.9% respectively, of all of the computer science doctorates earned in the U.S. There were no reported computer science doctorates for American Indian/Alaskan Native or Native Hawaiian/Pacific Islanders for 2010. Persons with disabilities are underrepresented in science and engineering despite earning more doctorates in S & E fields than non-S & E fields since 2007 (see <http://www.nsf.gov/statistics/wmpd/>).

Table 1. Number of ACW participants and number and percentage of survey respondents by ACW year.*

Year	Number of ACW Participants	Number of Survey Respondents by Year	Response Rate by ACW Year
2005	18	7	38.8%
2006	9	4	44.4%
2007	34	13	38.2%
2009	40	21	52.5%
2010	30	22	73.3%
2011	36	22	61.1%
2012	34	30	88.2%

*No ACW in 2008 to facilitate the transition to a spring workshop schedule.

Table 2. NSF submissions and award rate reported by ACW participants by submission year.

Year	Number of Proposals Submitted to NSF	Number of Proposals Funded by NSF	Percentage of Proposals Funded by NSF
2005	2	1	50%
2006	8	4	50%
2007	10	7	70%
2008	3	1	33%
2009	6	5	83%
2010	8	3	37%
2011	10*	1	10%*
Total for all Years	47	22	47%*

*Ten proposals were submitted for which a funding decision was given. An additional three proposals were submitted but a funding decision was not provided.

Longitudinal Evaluation

The purpose of the longitudinal evaluation was to ascertain the longer-term (two to seven years) effects of the workshop and the application of knowledge, skills, and attitudes attendees have demonstrated after participating. We sought to answer questions, such as:

- How did the participants apply the proposal writing components and to what extent has this yielded successful grant awards?
- How did learning about both the readily apparent and subtler aspects (unwritten rules) of promotion and tenure affect participants' understanding of the promotion and tenure process?

Methods. The evaluation consisted of a mixed-methods design that included focus groups and interviews conducted during the 2012 ACW and an online survey administered summer 2012 to ACW participants from 2005 through 2012.

Focus Groups and Interviews. Two focus groups (consisting of three past

participants each) and eight individual interviews were conducted at the 2012 ACW. The interviews and focus groups consisted of 14 participants: 11 men and three women. Six were African American/Black; seven were Hispanic/Latino—heritages included Venezuelan, Cuban, Peruvian, Columbian, Puerto Rican, and Mexican; one was East Indian and also deaf (using sign language). Two were Ph.D. students; three were associate professors (two are preparing their dossiers for promotion to full professor); one was a fifth-year assistant professor who submitted his promotion and tenure dossier in August 2012; three are assistant professors; one was a full professor (he was also a full professor when he first came to the workshop as a participant); two were researchers; and two were administrators.

Survey. The survey was administered online during May and June 2012. Of 166 ACW participants who attended from 2005–2012 for whom an

email address could be identified, 153 email messages with the survey link were delivered. Eighty-four ACW participants completed the survey, which is a response rate of 54.9%. Table 1 displays the number of ACW participants and the number and percent of survey respondents by ACW year.

The race/ethnicity of the ACW participants that responded to the survey is presented in the accompanying figure. The respondents were evenly split by gender: 41 (49%) female; 42 (51%) male. One individual did not respond to this item. Seven individuals reported having a disability (visual, hearing, mobility).

Results: Grant Proposal Submission and Awards

Given the importance of grant productivity to the success of tenure-track faculty, the ACW addressed this topic by structuring mock review panels that included the review of funded and unfunded proposals. The interviews and focus groups revealed unanimous agreement that this component was particularly helpful. Specifically, it helped “demystify” the process and informed participants of the proposal writing process (for example, how to frame the relevance and potential impact, how to find the best-fitting grant, contacting the program manager, opening one’s options to various agencies, and having others edit the proposal). One participant stated that he uses the workshop PowerPoint slides as a checklist when writing proposals. Comments included that the mock review panels increased self-efficacy in proposal writing, a greater propensity to write proposals as compared to be-

A unique feature of the ACW is that it includes mock review panels in which participants review proposals.

fore their participation in ACW, and receiving awards.

There are other resources available to assist with early career faculty and doctoral students in computer and information science and engineering with grant writing, professional development, and networking. A few to note include the NSF Computer and Information Science and Engineering Career (CISE) Workshops (<http://www.cis.temple.edu/NSFCareer2013/>) that were offered twice in 2012 with two workshops planned for 2013. The Richard Tapia Celebration of Diversity in Computing (<http://tapiacconference.org/2013/>) is an annual conference that provides professional development and networking opportunities focused on academic careers. The Computing Research Association's Committee on the Status of Women (CRA-W) (<http://cra-w.org/ArticleDetails/tabid/77/ArticleID/50/Career-Mentoring-Workshop-CMW.aspx>) and the Computing Research Association (<http://cra.org/events/career-mentoring/>) each organize a faculty-mentoring workshop in alternating years. The National Institute for Faculty Equity (NIFE) convenes a workshop at Georgia Tech for minorities in Engineering (<http://serc.carleton.edu/facultyequity/workshop12/index.html>).

A unique feature of the ACW is that it includes mock review panels in which participants review proposals (prior to attending) and engage in a discussion during the mock review process. The survey revealed 96% of the respondents indicated the mock review panel was essential or helpful to their gaining knowledge and skills related to grant proposal development. Fifty-two percent reported they utilize this information frequently in their career.

ACW participants were asked to list all full grant proposals submitted (as PI or Co-PI) since 2005 and the status. Funding agencies included NSF, Department of Education, Department of Defense, NIH, CDC, CISCO, and NSERC. The respondents reported 119 proposals submitted from 2005 to 2011 with 63 proposals funded. The percent of proposals funded was an impressive 53%.

The U.S. National Science Foundation (NSF) is where the largest number of proposals were submitted.

There is a significant focus at the ACW on navigating the promotion and tenure process.

Table 2 shows the NSF submissions and award rate reported by ACW participants by submission year. While causal inferences cannot be made between NSF grant awards and ACW participation, this table illustrates an impressive pattern of success in securing NSF grant awards. One workshop participant commented: "Before the workshop, I had read proposals but I had no idea of how to go about critiquing proposals or what to look for. After this workshop, I was able to know what to look for, what was good, and how to apply it to what is required..." Eighty-eight percent reported the ACW was essential or helpful in their acquiring the skills and knowledge to write successful funding proposals.

Results: Promotion and Tenure

There is a significant focus at the ACW on navigating the promotion and tenure process. This program component was also effectual, even for those who have chosen alternative careers. Participants used this information to guide their activities and decisions, be it obtaining more information from their institution on their tenure process or choosing an alternative career path. For one participant who had not previously considered academia, this component opened up that option for him. For another participant, it helped him in planning his path from doctoral student to professor. The following quote indicates how the knowledge gained was applied: "The first thing I did was going back to my personnel committee...and I learned the process of evaluation at my institution..."

One participant with a disability conveyed a similar perspective—that if one does not know what to ask or how

to seek out the information, then it is very possible that the individual is at a disadvantage and may never obtain it: "To me, I think, being deaf, I didn't feel my having the socialization outside of being here (at the ACW) because it enabled me to get that information and knowledge that I was missing. ... It helped me to ask the right questions. Sometimes you need the knowledge to know what questions to ask."

Fifty (63%) of the survey respondents were in full-time faculty positions. Most (27) were assistant professors; six were post-doctoral fellows. Almost all (90%) responded that the ACW Promotion and Tenure panel was essential or helpful in their gaining knowledge and skills in navigating this process. Ten respondents reported that since their first ACW, they have been promoted from assistant to associate professor. Eight doctoral students acquired assistant professor positions. The promotion and tenure component was critical for graduate students transitioning into faculty roles and current faculty progressing through the tenure process.

Conclusion

The findings from this evaluation should be of particular interest to department heads, dissertation directors, and those supervising postdoctoral fellows. The results emphasize that the ACW's unique configuration of peer mentoring, professional development, and mock proposal reviews is highly beneficial for early career faculty yielding immediate and sustained impact. 

References

1. Gates, A.Q., Ladner, R., Taylor, V., and York, B.W. Academic Career Workshops for underrepresented groups. *Computing Research News* 24, 4 (2012); http://cra.org/resources/crn-online-view/academic_career_workshops_for_underrepresented_groups/
2. NSF Survey of Earned Doctorates; <http://www.nsf.gov/statistics/srvydoctorates/>

Denise Ward Hood (dwhood@illinois.edu) is an assistant professor of Education Policy, Organization, and Leadership at the University of Illinois-Urbana.

Stafford Hood (slhood@illinois.edu) is the Sheila M. Miller Professor and Associate Dean for Research and Research Education, and the director of the Center for Culturally Responsive Evaluation and Assessment (CREA) at the University of Illinois-Urbana.

Dominica McBride (dmcbride@thehelpinstitute.org) is the co-founder and president of the The HELP Institute and the director of research at the Community Mental Health Council in Chicago, IL.

Copyright held by author.



The Profession of IT

Moods, Wicked Problems, and Learning

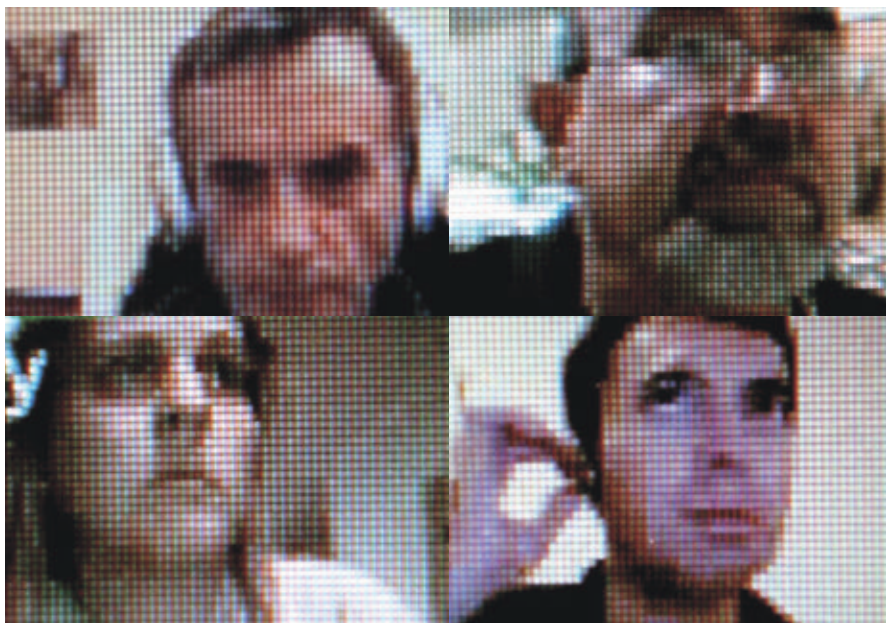
Wicked problems and learning environments present tough mood challenges for leaders and teachers. Telepresence and sensory gadgets are unlikely to replace physical presence in these areas.

IN MY PREVIOUS COLUMN (December 2012), I discussed the importance of understanding and being able to interact with different moods of individuals and groups. Positive moods enhance individual and team performance; negative moods detract and can render teams and groups dysfunctional. Skilled managers, facilitators, and teachers are keenly aware of moods and emotions. They know how to guide their teams through the moods necessary for a successful project, problem resolution, or learning.

There has recently been a lot of discussion about the role of technologies in two important areas: resolving wicked problems and learning. Managing moods is very important in these areas. I am very skeptical about telepresence proposals for these areas: they obscure moods.

Moods for Wicked Problems

Many modern social and technology challenges can be classified as wicked problems. A wicked problem is a very messy social tangle.¹ All prior attempts at solving it have failed. No one has enough power or resources to impose a solution, but everyone has enough power to block someone else's disagreeable proposals. The underlying difficulty is that the participants do not have a shared interpretation of the issue and they distrust each other; therefore, they cannot generate



a mood of solidarity to move in any direction. It may even compound their difficulties to call their issue a "problem," since they cannot even agree on a problem statement.

The term "wicked problem" is also used for problems in complex systems. The players agree on a problem statement, but cannot find a solution because extreme complexity hides it. Climate change modeling or finding effective new drugs look like wicked problems in this sense. With enough time and effort, we can find enough structure and recurrences in the system to solve these problems. The recent discovery of the Higgs boson at

the CERN facility in Switzerland culminated a search by 5,000 physicists spanning nearly 50 years.

I dislike calling complex-systems problems "wicked." I want to reserve the term for social tangles. No amount of scientific understanding will resolve social tangles; we have to address the core issue of the human disagreement and see if any agreement can be found.

Skilled facilitators have worked out processes that help the parties in a wicked problem find agreement; examples are the Appreciative Inquiry, Layton-Strauss, and Charrette processes.^{1,2} Trained facilitators help the parties find a shared interpretation and

develop action plans to move with it. The process can be described as a series of moods:

► *Appreciation.* Each player comes to appreciate all the points of view and concerns of the others. Some players modify their own concerns in the process. They develop a feeling that their concerns are understood and appreciated by the others.

► *Speculation.* The players cooperate on developing some possibilities for action, but do not commit to any particular action. After the possibilities are out in the open, they sort through to find out which ones take care of concerns in the group. This will help select a small set of promising actions.

► *Resolution and ambition:* The group commits to actions, usually performed by different teams tackling different aspects of the issue. The group sees the teams as parts of an experiment—try multiple actions and see which ones produce movement.

► *Follow up:* The group assigns managers to watch over the various action teams and see them through to completion. They agree to meet together again to renew their shared interpretation, evaluate previous actions, and commit to new actions.

The skilled facilitator can sense when the group has achieved each of these moods and only then moves on to the next stage. If the facilitator tries to push to the next stage too soon, the whole process may fall apart.

Moods for Teaching and Learning

Learning is not just a classroom activity; it is a team activity and professional responsibility. In the previous column, a table listed moods commonly encountered in the workplace. Seven of them directly affect learning—wonder, curiosity, inquiry, perplexity, apathy, and confusion. Learning from a mistake or breakdown is much harder if the person has a negative mood about it. The best moods are wonder and curiosity: the person knows there is something to learn, desires to learn, and embraces that something good will come from learning. The worst moods are apathy and confusion: the person is either indifferent or is annoyed at the mistake or breakdown and blames it on someone or something else. The teacher or manager seeks to transform

the mood of a confused person to inquiry, curiosity, or even wonder.

Managers who espouse “fail fast and often” are trying to predispose their people to accept failures and mistakes, inquire into what can be learned, and take new actions. They are trying to dispose their people toward wonder and inquiry, and away from confusion, procrastination, and resentment over wasted effort.

Recall situations where you did not know what to do, or you were surprised that an event did not go your way. How did you react? Do you have a conditioning toward confusion and away from wonder or inquiry? Do you procrastinate when you see that an inquiry might be useful?

Other moods come into play when it comes to learning to function effectively in a domain. Hubert Dreyfus discusses six stages of learning: beginner, advanced beginner, competent, proficient, expert, and master.³ Each stage marks a deeper level of embodiment of skill in the domain. The beginner has no embodied skill and performs solely by following the rules told by the teacher. The master relies completely on embodied skill and does not consciously apply rules when performing. The beginner is not attuned to the moods and emotions of people in the domain; the master is exquisitely attuned. Each stage has a characteristic mood that a teacher must foster. Only a person at a higher stage can be an effective teacher for a person at a lower stage. Viewed in this way, the learning process is a rich trove of moods.

Can It be Done by Telepresence?

It is interesting that so many common patterns of collaboration—one-on-one, teams, wicked problems, and teacher-student—all depend on individual and group moods. The skilled manager, leader, teacher, or facilitator must build on positive moods, rechannel negative moods, and remove from the team those who will not give up toxic moods.

Given that many teams and classrooms are now dispersed, it is important to ask how well can a leader perform these functions from a distance. Can telepresence replace physical presence? Can we deal effectively with emotional issues in the workplace via

Calendar of Events

March 16–17

ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments, Houston, TX, Sponsored: SIGPLAN, SIGOPS, Contact: Steve Muir, Email: steve@grimupnorth.org

March 18–22

Design, Automation and Test in Europe, Grenoble, France, Sponsored: SIGDA, Contact: Enrico Macii, Email: enrico.macii@polito.it

March 18–22

SAC 2013, Coimbra, Portugal, Sponsored: SIGAPP, Contact: Shim Yong-Sang, Email: yong_shim@sdstate.edu

March 19–22

International Conference on Intelligent User Interfaces, Santa Monica, CA, Sponsored: SIGCHI, SIGART, Contact: Jihie Kim, Email: jihie@isi.edu

March 20–24

Laval Virtual – International Conference on Virtual Reality and Converging Technologies, Laval, France, Contact: Matthieu Lepine, Email: mlepine@laval-virtual.org

March 22–24

Symposium on Interactive 3D Graphics and Games, Orlando, FL, Sponsored: SIGGRAPH, Contact: Marc Olano, Email: olano@umbc.edu

March 24–27

International Symposium on Physical Design, Stateline, NV, Sponsored: SIGDA, Contact: Cheng-Kok Koh, Email: chengkok@ecn.purdue.edu

March 24–29

Aspect Oriented Software Development, Fukuoka, Japan, Contact: Hidehiko Masuhara, Email: masuhara@graco.c.u-tokyo.ac.jp

email? Can we operate small teams with Skype videoconference? Can we manage larger projects with Cisco Telepresence? Can we facilitate a wicked problem group on the Internet? If not now, are there tools on the horizon that would permit any of these things in the future?

Let's consider some of the media available now. Email is very good when we have expository communications or simple coordination actions (requests, promises, deliveries, settlements). But it is notoriously bad for dealing with emotional people or situations. Email users are well advised to avoid responding to emotional email messages and instead to call or visit the other person.

Similarly, it is very difficult to conduct a speculation by email or other online venues such as a real-time wiki that only share written statements. Without presence, all the subtle cues of gestures, postures, voice tones, mood sensing, emotional reactions, and excitements are missing or difficult to gauge.

Some teachers have successfully used in-class "clicker" systems to get quick student feedback on comprehension questions. But experiments where meeting participants click "mood meters" to signal their moods have proved superficial because many participants do not understand their own moods, and because linguistic indicators are not complete characterizations of moods. When the group members already know each other well, a videoconference can work because the team members have already developed a background of trust.

In the 1960s, Paul Eckman created a facial action coding system (FACS). That system has been perfected over the years and has married with modern vision processing to give us sophisticated technology for inferring people's emotions from facial expressions. Market researchers use FACS software to discover how people are reacting to ads (see *Emotionomics*, by Dan Hill, Kogan Page Publisher, 2010). It is not hard to imagine a system in the near future that individually tracks the facial expressions of everyone in a meeting and provides the facilitator with a display showing the kinds of emotions in the room and giving display markers that zero in on faces with

Technologists who believe they can replace facilitators and teachers with machines are mistaken.

particular emotions. Still, such a sophisticated system is unlikely to duplicate the skilled facilitator. It provides its information only through a visual channel. It cannot provide information of the richness sensed in the body by the facilitator moving around the room among the people.

Hubert Dreyfus, noted earlier, devotes a whole chapter to the issue of telepresence in teaching. His question is: How far up on the learning scale from beginner to master can a student progress when the only contact with the teacher is via automated courseware and telepresence? Telepresence would include real-time voice and video interaction between student and teacher, video feeds that permit students to see what the teacher sees and vice versa, and tight integration with display tools such as presentations, pictures, images, and sounds. It would also require technology that supports two-way eye contact—generating the mutual feeling that the other person is looking back at you and is present with you. From an examination of teachers in classrooms, Dreyfus concludes that teachers are exquisitely sensitive to the moods in the room. How do teachers tell when students are generally receptive to a topic or discussion? When they are engaged? That a student's question resonates with the whole class? That a quiet student has a burning question? Experiments with special tracking devices that permit an observer to see exactly what the teacher sees indicate observers cannot sense the moods the teacher is responding to.

For all these reasons, Dreyfus concludes that today's technologies are barely able to allow a telepresent teacher to guide a student up to the level of

competence. Dreyfus is very skeptical that we will figure out how to do the higher stages via telepresence. The human body's ultrasensitive ways of detecting and responding to moods and emotions are not likely to be simulated by machines anytime soon.

Still, the future is full of surprises. The picture may be brighter for a judicious combination of telepresence and physical presence. MOOCs—massive open online courses—represent a new generation of courseware now making college courses available for free. The organization of material and production quality makes them better than many existing courses. They permit much better interactivity with the teacher than in a 500-student amphitheater. Students form their own group "meets" in local Internet cafes so that they can physically study together. Their instincts to meet overcome the limitations of the Internet by fostering the positive moods of learning. If a local study group includes an experienced coach, the students might be able to move up the learning ladder effectively. Dreyfus may be right that "pure" telepresence cannot do the job, but coached hybrids might.

Conclusion

Two contemporary challenge areas commanding a lot of attention—wicked problems and education—are approachable by leaders and facilitators skilled at reading moods and guiding others to the moods needed to reach their goals. Technologists who believe they can replace facilitators and teachers with machines—such as pure telepresence and social media gadgets—are mistaken. The human body is exquisitely sensitive to subtle signals that enable it to read moods. No one knows how to sense, transmit, or receive these signals. In these areas, humans must remain in the loop. **C**

References

1. Denning, P. and Dunham, R. *The Innovator's Way*. MIT Press, 2010.
2. Denning, P. and Yaholkovsky, P. Getting to we. *Commun. ACM* 51, 4 (Apr. 2008), 19–24.
3. Dreyfus, H. *On the Internet*. Routledge, 2001.

Peter J. Denning (pjd@nps.edu) is Distinguished Professor of Computer Science and Director of the Cebrowski Institute for information innovation at the Naval Postgraduate School in Monterey, CA, is Editor of *ACM Ubiquity*, and is a past president of ACM.

Copyright held by author.

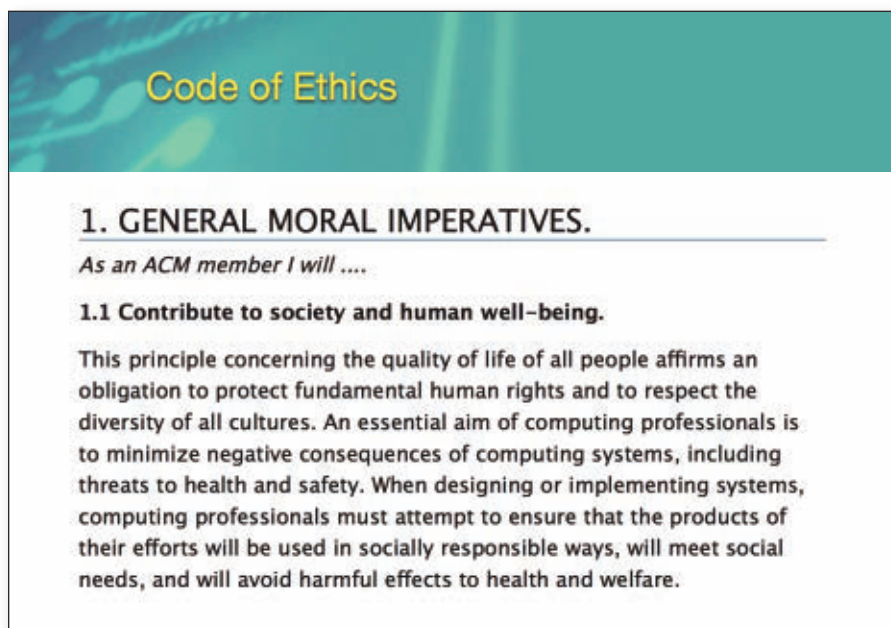
Computing Ethics Ethics Viewpoints Efficacies

Seeking answers to ethical concerns.

IN 2008, I ACCEPTED the ACM editors' invitation to solicit manuscripts for *Communications'* Computing Ethics column. This is the tenth installment published since the inception of the column, which has featured a variety of authors covering a wide range of ethical issues—from ethics and robotics in civilian and military applications, to ethics and smart grid technology, to the question of whether software engineering qualifies as engineering. But I have not heard much from the readers of the column, which leads me to wonder whether these questions are those with salience to ACM members.

One approach to this question was to examine the ACM website. There, the members who were selected for profiles on the home page indicated for me an emphasis in the organization on achievement in industry. This emphasis prompts me to ask whether members might be less interested in the ethics of research and development and more interested in ethics questions that arise in practice.

This distinction does not seem so important looking at the section of the ACM website devoted to the association's history. The history indicates rapid post World War II growth from its founding as the Eastern Association for Computing Machinery in fall, 1947; dropping "Eastern" from the name in 1948; and instituting a constitution in 1949. In the initial meeting, the publicity stated that the purpose of the association "would be to advance the



The image shows a screenshot of the ACM Code of Ethics page. It has a teal header with the title "Code of Ethics" in yellow. Below the header, the text is white on a dark background. The section is titled "1. GENERAL MORAL IMPERATIVES." and includes a sub-section "1.1 Contribute to society and human well-being." The text describes the principle concerning the quality of life of all people and the obligation to protect fundamental human rights.

Code of Ethics

1. GENERAL MORAL IMPERATIVES.

As an ACM member I will

1.1 Contribute to society and human well-being.

This principle concerning the quality of life of all people affirms an obligation to protect fundamental human rights and to respect the diversity of all cultures. An essential aim of computing professionals is to minimize negative consequences of computing systems, including threats to health and safety. When designing or implementing systems, computing professionals must attempt to ensure that the products of their efforts will be used in socially responsible ways, will meet social needs, and will avoid harmful effects to health and welfare.

Excerpt from the ACM Code of Ethics.

science, development, construction, and application of the new machinery for computing, reasoning, and other handling of information." Presently, as the website indicates, the ACM constitution summarizes its purview as that of "an international scientific and educational organization dedicated to advancing the art, science, engineering, and application of information technology, serving both professional and public interests by fostering the open interchange of information and by promoting the highest professional and ethical standards."

This information is found in the history section of the site, which has

less visibility than does the home page itself, and it is likely that site visitors view the history section distinctly less often than they do other sections of the site. It is heartening to see the recognition of ethics in the last version of the constitution. Nonetheless, the emphasis on IT advancement indicates perhaps where the priorities of ACM members are. This is neither surprising nor inappropriate; yet, it says to me that the heart of the organization and its members lies with innovation itself, rather than the ethics of innovation.

It also suggests to me that recognizing this fondness may be inseparable from finding examples and issues of

ethics that resonate with the members. Meeting this criterion means the discussion needs to raise concerns that arise in daily practice or at least over the course of a project. Perhaps they need to be more realistic than speculative and capture the different interests at stake in the activity and its resolution. Sometimes they could be quite brief and provide a specific piece of advice; other times they would raise a series of issues and challenges that will take time to play out and resolve. Are these appropriate criteria? Do they capture members' priorities?

The Center for Engineering, Ethics, and Society at the National Academy of Engineering, which I direct, manages the Online Ethics Center (<http://onlineethics.org>). It has a large selection of cases, many with ethics commentaries, and I believe that more than a few of them speak to the issues that IT scientists and engineers face often. I am going to use this column to point readers to a few, and ask you whether these cases and commentaries, where there are comments, are of significance to you and portray phenomena that you recognize as those you or colleagues face. If you have other cases (suitably anonymized) that you wish to share, the OEC will consider publishing them and, as possible, finding people to comment on them. You can contact me with ideas, cases, or other relevant information at rhollander@nae.edu.

Online Ethics Center cases take numerous forms. Some are quite detailed, with sections of description as well as elaborate commentary. Some are quite short with brief commentaries. A longstanding, extensive case is that of "The Killer Robot" by Richard G. Epstein, which is the tale of software gone rogue in a medical application. See <http://www.onlineethics.org/Resources/Cases/killer-robot.aspx>. A more recent short case involves a difficulty an engineer was having figuring out his obligations to former and new employers; called "Obligation to Client or Employer," this case is not specific to IT engineering but it raises questions pertinent to IT engineers and the commentaries provide good advice; see <http://onlineethics.org/Resources/Cases/Obligation.aspx>.

Having things turn out okay does not entail the conclusion that the actions taken were ethically appropriate.

The OEC recently posted a case titled "Occidental Engineering." I think it has a lot to offer, both because it describes a not infrequent kind of problem and demonstrates what I believe is an important aspect of ethics—that having things turn out okay does not entail the conclusion that the actions taken were ethically appropriate. It also contains useful commentary from the author about the ethical dimensions of the case and how to teach it. One thing it does not have is a set of quantitative problems that might be relevant to engineering students considering the case; if readers have relevant suggestions for, or can develop, problem sets for this or similar problems that could be posted, please let me know.

Here is a brief summary of the case: A software engineer in the aerospace division of Occidental Engineering is working on a contract from the FAA that the company had "lowballed" in order to beat out competitors and garner much needed business. But the company therefore had to underfund and understaff the project. A version of the prototype (for a next generation air traffic control system) needs to be delivered, fully certified for system integration and test, in three days, but that does not leave enough time for the engineer and his team to resolve a little problem before the delivery date—the problem being that when there are too many aircraft in the system, it will sometimes lose track of one or more of them. The team has traced the problem to a subtle error in memory allocation and reuse. They are confident that they can fix it, but it will take a month or more.

The government has developed a new, get-tough policy on missed dead-

lines and cost overruns, and Occidental is afraid that if they miss this deadline, they would be fined and lose the remainder of the prototype contract; and they might not be allowed to bid on the contract for the full system resulting in thousands of lost jobs. They consider and reject the idea of a quick patch. Their management decides they should deliver the software as-is, noting that FAA testing plans will include an active backup system when they do get to live tests, and they will do those only at a small airport. They will not overload the system. After that they request changes, and even if not, the company can provide an updated version of the program with the bug fix. If they see the problem, the company can claim it was a random occurrence. The important thing is no one is in any danger, so the system can be certified as safe, for the use to which it will be put.

In the end the engineer signs off; the testing works out okay; the problem is solved; and the company gets much needed business. The lead engineer takes early retirement once the prototype project is finished, in order to write a book on software testing. He feels the book should have a chapter on ethics, but he can never bring himself to write it.

The questions the author asks at the end of the case are, "What do you think about the engineer's decision? Was it ethical?" How would you answer, and what are your justifications for your answers, particularly your answer to the second question? Think about your answers, and then you might want to visit the site and review the case and commentary. Michael McFarland, S.J., a computer scientist and the former president of College of the Holy Cross, was a visiting scholar at the Markkula Ethics Center. He wrote the case and posted it to that center's site in June, 2012. He provides an extensive ethics tutorial with it. You can find the case and commentary, with a link to the Markkula site, at <http://onlineethics.org/Resources/Cases/OccidentalEng.aspx>. 

Rachelle Hollander (rhollander@nae.edu) is the director of the Center for Engineering, Ethics, and Society at the U.S. National Academy of Engineering in Washington, D.C. She is a member of the Governing Board of the Association for Practical and Professional Ethics.

Copyright held by author.

Viewpoint

Can Computer Professionals and Digital Technology Engineers Help Reduce Gun Violence?

Ten idea seeds.

REGARDLESS OF ONE'S position on gun control, few of us consider it tolerable for mass-shootings like the recent ones in the Sandy Hook Elementary School, or in a Colorado movie theater, or in an Oregon shopping mall, or on a Texas Army base, or in a Norwegian youth camp, to continue to occur. The question is, what can we do about it? We can of course engage as citizens in political debates, political actions, and policymaking on gun legislation and policy. But I would like to ask if there is any way that ACM members, as computer scientists, engineers, and digital technology experts, can use our technical expertise to reduce the frequency and casualty count of shooting crimes.

This Viewpoint is an attempt to start us thinking about that.

Eventually Everything Goes Digital

In the past several decades, many devices that were mechanical or analog electrical 50 years ago became digital: calculators, cash registers, watches, phones, cameras and photographs, movies, music players, musical instruments, washing machines, sewing machines, power supplies, roledexes, copiers, calendars, televisions, and others. As the trend continues, more objects and appliances are becoming digital: books, magazines, newspapers, metronomes, shavers, toasters, cof-



feemakers, cars, planes, trains, lighting systems, batteries, power grids, homes, and others.

When an object or appliance goes digital, capabilities emerge that the old pre-digital versions did not have. The object is transformed. Some devices, after undergoing a digital transformation, in turn transform our behavior and in some cases our society.

When Guns Go Digital, What Becomes Possible?

Ignoring high-tech military weapons like Stinger missile launchers, guns are still stuck in the analog, physical world. What if, like so many devices and appliances, guns became digital? What might be possible?

In particular, as guns become digital—and they will, like so many devices



ACM Conference Proceedings Now Available via Print-on-Demand!

Did you know that you can now order many popular ACM conference proceedings via print-on-demand?

Institutions, libraries and individuals can choose from more than 100 titles on a continually updated list through Amazon, Barnes & Noble, Baker & Taylor, Ingram and NACSCORP: CHI, KDD, Multimedia, SIGIR, SIGCOMM, SIGCSE, SIGMOD/PODS, and many more.

For available titles and ordering info, visit:
librarians.acm.org/pod

before them—how can we as responsible computer and digital technology professionals help ensure they are safer and less useful in crimes?

Ten Idea Seeds: If Guns Were Digital...

I will seed our thinking with some blue-sky ideas generating from brainstorming. The ideas get wilder as you read down the list, but none are magical; all are based on technology that exists today in some form. In the spirit of brainstorming, please suspend disbelief, suppress your internal critic, and read on with an open mind. Then engage your creativity and critical thinking to improve on or replace these idea seeds.

1. *What if switching a gun out of safety mode (in which it will not fire) required a combination or key?* Actually, many guns already have this. What if *all* guns did? What if unlocking the safety required the holder to say a code? What if guns had a secret second safety in addition to the primary safety? What if the secret safety were invisible (that is, internally located) and digital, requiring touching the gun to an unlocking device, like the anti-theft tags on many retail products that are deactivated at checkout? What if unlocking the safety required simultaneous operation by two people, one with the gun and one elsewhere, just as arming a nuclear missile requires at least two people working separately? What if a gun's safety mode was not controlled by a switch that stays in the ON position, but rather required a continuous data-feed from somewhere to remain ON, for example, the gun owner's re-

peated signals of consent or his or her vital signs?

2. *What if operating a gun required authentication, so it would function only for an authorized user?* What if gun owners had to login to their guns or otherwise (for example, biometrically) identify themselves to their guns, before the gun would fire or perform any function? This would make stealing guns useless. It would prevent a person's own guns from being used against them or other persons. For example, if Nancy Lanza's guns would work only for her, not for her son Adam, it would have been more difficult for him to shoot her and then carry out the mass shootings at Sandy Hook Elementary School.

3. *What if guns could raise alarms if stolen?* Small "bugs" are available today that can be attached to objects or people and raise an alert if removed from a specified zone. Such "bugs" can already be attached to guns. What if there were special gun "bugs" that could be programmed to permanently disable the gun if it were not returned to the home zone within a short grace period?

4. *What if every gun could be easily and accurately located?* Consumers today can buy key-fobs and stick-on tags that can be attached to often-lost items (such as keys, glasses, pets, children) and, on demand, emit sounds or radio signals that allow them to be found quickly (for example, ClickNDig, EZ-Find, Loc8tor). What if each consumer gun, when signaled, would emit a sound or signal that allowed the gun's location to be pinpointed? People attempting to carry such guns through security checkpoints could be easily spotted without the need for intrusive pat-downs.

5. *What if a gun would not fire if it detected alcohol in the breath of its holder?* Some of today's cars "sniff" the driver for alcohol breath and lock the steering wheel or will not start if they detect it. Guns could do something similar. Even gun-rights advocates admit that drunkenness is a significant factor in many shooting crimes.

6. *What if all guns inside a specified zone could be blocked from firing?* Cellular phone service in an area or a building can be jammed or blocked, rendering cellphones in that area use-

**What if, like
so many devices
and appliances,
guns became
digital? What might
be possible?**

less. Movie theaters, shopping malls, and schools, for example, could block guns from firing on their premises. What if all guns had a chip that classified them as “consumer,” “law enforcement,” “military,” and so forth. Law enforcement officials entering an area where a shooter was at large could temporarily render all nearby guns, or all “consumer” guns, inoperative. What if the firing mechanism in guns could be fused by a high-powered electronic pulse, just as such pulses can fuse critical parts in car engines and electric motors?

7. *What if each individual gun could be shut off or rendered inoperative remotely?* Smart phones and laptop computers can be tracked (for example, Apple’s “Find my iPhone” service) and even shut down remotely if they are on line. For decades, people have been able to open and lock their cars from a short distance away using remote control key-fobs. Many new cars have digital monitor-and-control boxes connected to GPS and cellular services that allow car owners to remotely monitor their car’s location and speed and even shut them down remotely (for example, Autonet Mobile, Mavison, OnStar, and TiWi). Some of today’s homes allow their residents to monitor or control certain household functions by phone or Internet when away. What if every gun contained a cellular chip that allowed the gun’s owner (or law enforcement if the gun owner delegates control to them) to block the gun from firing? What if guns could be jammed by purposefully infecting them with a digital virus or worm?

8. *What if guns would not fire if aimed at a person?* Many new cameras have limited face-recognition capability to help focus the camera on the subject’s face. Could consumer gun sights use similar technology to engage the safety lock when aimed at a person rather than a soda can, bird, or bear? Could guns provide auto-safety settings to allow their owners to specify what types of targets are valid or invalid?

9. *What if a gun could refuse to fire if something “felt wrong” about how it was being used?* Some military weapons, such as surface-to-air missile batteries, already contain artificial intelligence software that uses sen-

Engage your creativity and critical thinking to improve on or replace these idea seeds.

sors and human input to build and analyze complex situation profiles to decide quickly whether approaching objects are threats or not (for example, Israel’s “Iron Dome” missile defense system). Some military weapons systems can actually abort if their analysis suggests firing was in error or is likely to cause unintended casualties or damage, and other safety-critical systems can self-monitor and take action to prevent accidents or limit damage (see Nancy Leveson’s books, *Safeware* and *Engineering a Safer World*). Could similar technology be made less expensive and built into guns? Could future guns—let’s call them M160s, AR-1500s, and AK-47000s—be programmed to shut down if their sensory analysis suggested they were being handled by a child or used in a crime?

10. *What if shooting were digitized?* What if digital guns only operated in the “cloud,” that is, in virtual worlds? What if arguments that escalated to violence could take place only in the Metaverse (see Neil Stephenson’s novel *Snow Crash*) or the Matrix (see the movie by the Wachowski brothers)? What if the Second Amendment to the U.S. Constitution were interpreted to apply only online, in game-like environments: in cyberspace you can have all the guns you want and do all the shooting you want, but in meatspace there are serious restrictions—perhaps no guns at all. What if nations, if they failed to resolve their conflicts in negotiations, fought things out only in *World of Warcraft* cyberscapes, rather than in real cities and landscapes where people, including civilians, suffer and die and valuable property is destroyed?

Conclusion

These ideas may seem like wishful thinking; they may seem impossible; they may even seem crazy or undesirable. Indeed, they may *be* impossible, crazy, or undesirable. But they are only “seeds” I am planting to start creative minds in the computer and digital products industry thinking, to see if there is any way we as computer and digital technology professionals can use our technical expertise to help reduce the ever-rising casualty count. Furthermore, since submitting the original version of this Viewpoint, I learned that others are thinking along similar lines: Jeremy Shane, a former U.S. Justice Department official during the George H.W. Bush administration, wrote an article for CNN: “Make Guns Smart” (<http://www.cnn.com/2013/01/09/opinion/shane-smarter-guns/index.html>) that offers similar suggestions.

Of course, applying our digital technology expertise to the problem of gun violence does not preclude us from also engaging as citizens in debate and political action. Gun violence, especially in the U.S., is a highly charged issue that must be addressed in many different ways. I am engaged politically on this issue, and I hope you are too or in the aftermath of the shootings at Sandy Hook Elementary School will become so.

But I also hope that now you have digested the seed ideas I have listed, you can engage your critical mind and your creativity and help to grow the seeds into viable ideas or replace them with better ideas.

Let’s not just wait for others to solve the problem of gun violence in our society. It is too important. C

Jeff Johnson (jjohnson@uiwizards.com) is a principal at two usability consulting firms: UI Wizards, Inc. and Wiser Usability, Inc. He was formerly Chair of Computer Professionals for Social Responsibility and currently serves on ACM SIGCHI’s U.S. Public Policy Committee. He is the author of *GUI Bloopers*, *GUI Bloopers 2.0*, *Web Bloopers*, *Designing With the Mind in Mind*, and (co-authored with Austin Henderson) *Conceptual Models*.

Copyright held by author.

Viewpoint

Funding Successful Research

A proposal for result-based funding for research projects.

RESearch foundations want to fund great research projects. However, a while back Bertrand Meyer wrote an interesting blog post: “Long Live Incremental Research.”¹ With examples he showed that many of the greatest research results could not possibly be projected in great sounding project descriptions. His conclusion is that we should drop the high-flying ambitions from research project descriptions, and instead support more incremental research proposals, hoping that great stuff will happen on the way. Indeed incremental research is perfect for research projects with predictable deliverables. However, I suggest an alternative conclusion: for some funding, we should drop the project description entirely.

Instead, we should initiate some pure result-based funding. An x-year grant could be based on results from the last x years. From a research foundation perspective, this eliminates the issue of unpredictable research, for this funding is not given for a projected future that may or may not happen. Rather it is rewarded for results already delivered. The researcher can at his own risk follow the craziest inspiration, but he or she has a strong incentive to make it work if he or she wants to secure result-based funding in the future. Result-based funding would only be applicable for researchers with a history of success, with emphasis on the more recent past, and the funding would only work for basic expenses



that are independent of the concrete project. In the U.S., for example, a baseline might be one or two months of summer salary and one or two graduate students. Junior faculty hired based on an impressive recent track record would be fully eligible. Senior faculty would need to demonstrate that they are still going strong. The simple point is to drop the project description and just reward what is already done.

Consider a researcher with a history of brilliant ideas taking research in surprising new directions. If we try casting this as a project, the referees will rightly complain: “It is not clear how the applicant will come up with a brilliant idea, nor is it clear what the surprise will be.” With such lack of focus and feasibility, a low project score is expected, and then the overall score

will be too low for funding, regardless of the researcher’s established record of success. However, research needs great new ideas. Therefore, we need some result-based funding so that we can support creative researchers with a proven talent for great new ideas even if we do not know how it will happen.

The aforementioned issue is often very real in my field of theoretical computer science. Like in other fields, theoretical research is only interesting if it contains surprises (otherwise it is more like development). A project plan would make sense if the starting point was a surprising idea or approach that it would take years to develop, but in theory, the most exciting ideas are often strikingly simple. When you first have such an idea, you are typically close to done, ready to start writing a paper. Thus, if you have the right idea when you apply for a grant, you will typically be done long before you get the grant. The essence of the research is an unpredictable search for powerful ideas and insights. Thousands of wild ideas may be tried in the search of a brilliant one that works. The most appropriate project description is just a description of the importance of the area to be researched and the type of results aimed for. The track record shows which researchers have the talent to succeed.

The problem (which may be much bigger in the EU than in the U.S.) for such dynamic research is when proposals are selected by project-oriented researchers who want structured

methodological plans, specifying how to attain the proposed goals, and who do not appreciate that a successful outcome depends heavily on the talent of the involved researchers. The philosophical difference is if we only count the creativity and originality specified up front in the project description, or if a researcher's demonstrated talent for creativity and originality is counted as an integral part of the research to be performed in the project.

Dropping the project description will greatly increase methodological diversity, allowing researchers to use the strategy that has proved most suitable for their area and their own talent and skills. As a simple example, Meyer suggested funding incremental research, hoping that great surprising things would turn up on the way. I favor the opposite strategy, spending as much time as possible pursuing overly ambitious targets, but being flexible about the results. Even if the high-flying targets fail, you do not need to come home empty-handed, for by studying the unknown you may discover something new, sometimes more interesting than the original target. From the perspective of ambition, I see it as an advantage to minimize time spent on easy targets, but foundations seem to prefer that you take a planned path with some guaranteed targets on the way. The point here is not to argue whether one strategy is superior to the other, but rather to embrace the diversity of strategies that work depending on the area and the individual researcher.

Perhaps more seriously, if a target is difficult to achieve, it may be because it requires an atypical approach that would not look reasonable to anyone else, but which may work for a researcher thanks to his special talents and intuition. Indeed, I have often been positively surprised seeing how others succeeded using an approach I had myself dismissed. As a project, such unbelievable approaches would fail on perceived feasibility, but the point in result-based funding is that researchers are free to use whatever approach they find most efficient. Funding is given to those who prove successful. This gives the perfect incentive to do great work; namely to secure future result-based funding.

Result-based funding would also reduce resources needed to evaluate applications. It is very difficult for a general panel to evaluate the methodology and the probability of success of a project. Moreover, it requires an intimate knowledge of a field to evaluate how big a difference a result would make relative to what is already known in the field. However, handling published results, we know what happened and if published in a strong venue, the experts have already verified the novelty to the field.

Some prestigious grants say they welcome high-risk high-gain research. Surprising breakthroughs in an important area would fall well within this scope. Having researchers with proven skills explore the area and follow their inspiration may be the optimal strategy, a bit like sending an expedition into an unknown territory. Uncertainty about what they would find should be no worse than high risk. In fact, based on past performance, it may be safe to assume they will discover something interesting, if not ground-breaking. However, when a project is scored based on focus and feasibility, projects where the end results are not predictable in advance will fail even if their expected return is very high. It has to be possible to get a high overall score for promising research even if it would not score well under standard project parameters like focus and feasibility. At the end of the day, what we want are results, not project descriptions, so what should determine the overall score is which proposal is expected to yield the greatest results.

The issue boils down to the formula used to compute the overall score of a proposal, the problem being when the score is based on a predefined weighted average, diluting the impact of any unique aspect. As a concrete case, I experienced an integration grant giving the established quality of the researcher a predefined weight of 30% of the total score. The remaining 70% of the weight was all about the projected future: project description (30%), implementation (20%), and impact (20%). The world's best most original researcher with the biggest prizes to his or her name can get at most 100% on established quality, contributing 30% to the total average. A more typical researcher may get 80% on established

quality, contributing $30\% \times 80\% = 24\%$ to the total average. The incremental advantage of the super-genius over the more typical researcher is thus a mere 6%, which is easily lost in the 70% of the weight devoted to the projected future. As a kind of entertaining example from the projected future, one question was: "Outline the capacity for transferring the knowledge previously acquired to the host." As a theoretician I thought the answer was simple: "The knowledge sits in my head so the transfer is complete on arrival. From my head, I will transfer knowledge and ideas to students, colleagues, and visitors." Naively I thought I would get 100% on this one, but my answer was deemed unconvincing, that is, 0%. The point I try to make here is not whether my answer was good or bad. My point is that while this transfer of knowledge may be critical in some cases, it is typically not an issue in my theoretical field. The general point is that the more standard parameters you involve in an average score, the more you favor standard proposals that these parameters apply to. However, what makes research special is normally something unique, for example, a great researcher, or a great idea for a project. To let the uniqueness come through, one should not average, but rather look at a maximum, possibly with a fail/pass on other parameters, allowing for some to be not applicable. The proposed result-based funding would cover the case of great researchers.

I have proposed the initiation of some pure result-based funding as a simple efficient method for basic support of successful researchers, giving them the freedom and incentive to seek great results even when these are not projectable. Project-based proposals would still be needed in many cases, for example, to justify expensive experiments. Because result-based funding is simpler to handle, it could be used efficiently as a first line of funding with smaller individual grants. 

Reference

1. Meyer, B. Long live incremental research. BLOG@CACM, June 13, 2011.

Mikkel Thorup (mikkel2thorup@gmail.com) is a professor in the Department of Computer Science at the University of Copenhagen in Denmark.

Copyright held by author.

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

Racing to unleash the full potential of big data with the latest statistical and machine-learning techniques.

BY ARUN KUMAR, FENG NIU, AND CHRISTOPHER RÉ

Hazy: Making It Easier to Build and Maintain Big-Data Analytics

THE RISE OF big data presents both big opportunities and big challenges to domains ranging from enterprises to sciences. The opportunities include better-informed business decisions, more efficient supply-chain management and resource allocation, more effective targeting of products and advertisements, better ways to “organize the world’s information,” and faster turnaround of scientific discoveries, among others.

The challenges are also tremendous. For one, more data comes in diverse forms: such as text, audio, video, OCR, and sensor data. While existing data management systems predominantly assume that data has rigid, precise semantics, increasingly more data (albeit valuable) contains imprecision or inconsistency. For another, the proliferation of ever-evolving algorithms to gain insights from data (in the name of machine learning, data mining, statistical analysis, and so on) can often be daunting to a developer with a particular dataset and specific goals: the developer not only has to keep up with the state of the art, but also must expend significant development effort in experimenting with different algorithms.

Many state-of-the-art approaches to both of these challenges are largely statistical and combine rich databases with software driven by statistical analysis and machine learning. Examples include Google’s Knowledge Graph, Apple’s Siri, IBM’s “Jeopardy!”-winning Watson system, and the recommendation systems of Amazon and Netflix. The success of these big-data analytics-driven systems, also known as *trained systems*, has captured the public imagination, and there is excitement in bringing such capabilities to other verticals such as enterprises, health care, sciences, and government. The complexity of such systems, however, means that building them is very challenging, even for Ph.D.-level computer scientists. If such systems are to have truly broad impact, building and maintaining them needs to become substantially easier, so that they can be turned into commodities that can be easily applied to different domains. Most of the research emphasis so far has been on individual algorithms for specific machine-learning tasks.

In contrast, the Hazy project (<http://hazy.cs.wisc.edu>) takes a systems approach with the hypothesis: The next breakthrough in data analysis may not be in individual algorithms, but in the ability to rapidly combine, deploy,



and maintain existing algorithms. Toward that goal, Hazy's research has focused on identifying and validating two broad categories of "common patterns" (also known as *abstractions*) in building trained systems (see Figure 1): programming abstractions and infrastructure abstractions. Identifying, optimizing, and supporting such abstractions as primitives could make trained systems substantially easier to build. This can bring us a step closer to unleashing the full potential of big-data analytics in various domains.

Programming abstractions. To ensure that a trained-system platform is accessible to many developers, the programming interface must be small and composable to enhance productivity and enable developers to try many algorithms; the ability to integrate diverse data resources and formats requires the data model of the programming interface to be versatile. A combination of the relational data model and a probabilistic rule-based language such as Markov logic satisfies these criteria. Using this combination, we have developed several knowledge-based construction systems (namely, DeepDive, GeoDeepDive, and AncientText). Furthermore,

our (open source) software stack has been downloaded thousands of times and used by different communities such as natural language processing, chemistry, and biostatistics.

Infrastructure abstractions. To build a trained-system platform that can accommodate many different algorithms and that scales to large volumes of data, it is crucial to find the invariants in applying individual algorithms, to have a clean interface between algorithms and systems, and to have a scalable data-management and memory-management subsystem. Using these principles, we developed a prototype system called Bismarck,¹⁰ which leverages the observation that many statistical-analysis algorithms behave as a user-defined aggregate in an RDBMS. The Bismarck approach to data analysis is resonated by commercial systems providers such as Oracle and EMC Greenplum. In addition, such infrastructure-level abstractions allow us to explore generic techniques for improving the scalability and efficiency of many algorithms.

Example Application: GeoDeepDive

An application called GeoDeepDive (<http://hazy.cs.wisc.edu/geodeepdive>) illustrates the Hazy approach to building trained systems. GeoDeepDive is a demo project involving collaboration with geology researchers to perform deep linguistic and statistical analysis over a corpus of tens of thousands of journal papers in geology. The goal is to extract useful information from this corpus and organize it in a way that facilitates geologists' research. The current version of GeoDeepDive extracts mentions of rock formations,

tries to assign various types of attributes to these formation mentions (for example, location, time interval, carbon measurements), and then organizes the extractions and documents in spatial and temporal dimensions for geoscientists. Figure 2 shows a high-level overview of how Hazy built GeoDeepDive.

Using the Hazy approach, the GeoDeepDive's development pipeline consists of the following steps:

1. The developer assembles data resources that are potentially useful for GeoDeepDive.
2. The developer composes feature-extraction functions that convert the data resources into relational signals.
3. The developer specifies correlations and constraints over the relational signals in the form of probabilistic rules; Hazy's infrastructure performs scalable statistical learning and inference automatically.
4. Hazy outputs probabilistic predictions for GeoDeepDive.

Input data sources. Hazy embraces all data sources that can be useful for an application. GeoDeepDive uses the Macrostrat taxonomy (<http://macrostrat.org/>) because it provides the set of entities of interest, as well as domain-specific constraints (for example, a formation can be associated only with certain time intervals). Google search results are used to map location mentions to their canonical names and then to latitude-longitude (lat-lng) coordinates using Freebase (<http://freebase.com>). These coordinates can be used to perform geographical matching against the formations' canonical locations (lat-lng polygons in Macrostrat). There are also (manual)

Figure 1. Hazy's programming abstractions and infrastructure abstractions.

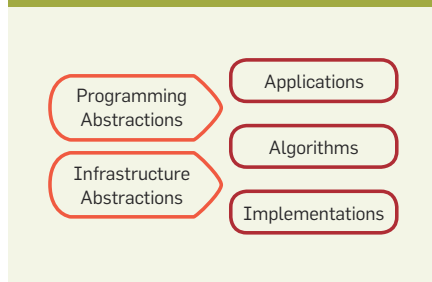
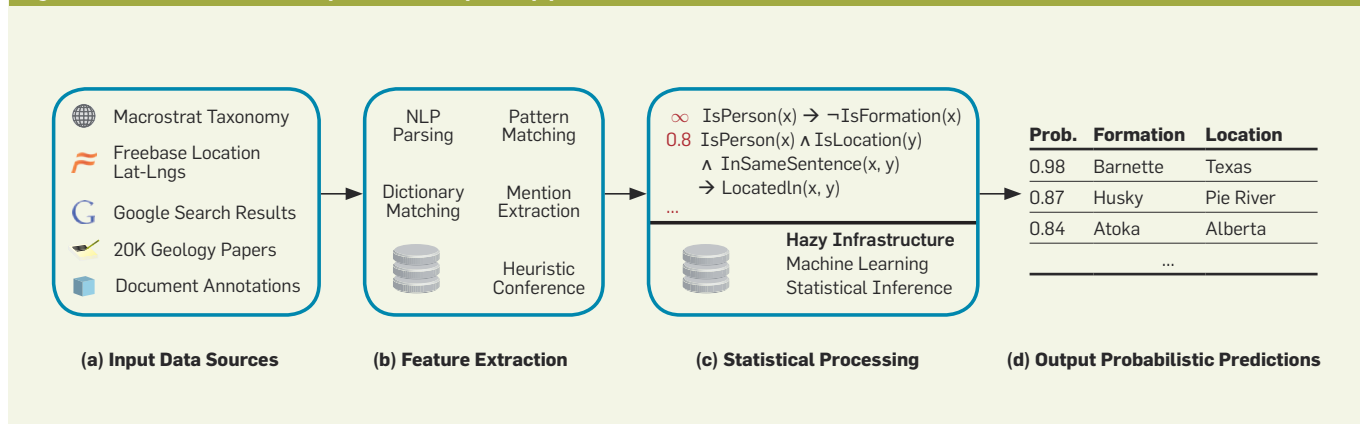


Figure 2. An overview of GeoDeepDive's development pipeline.



document annotations of textual mentions of formation measurements that serve as training data.

Feature extraction. The input data sources may not have the desired format or semantics to be used directly as *signals* (or *features*) for statistical inference or learning. The feature-extraction step performs such conversions. The developer specifies the schema of all relations, provides individual extractors, and then specifies how these extractors are composed together. For example, we (the developers) perform NLP parsing on the input corpus to produce per-sentence structured data such as part-of-speech tags and dependency paths. We then use the Macrostrat taxonomy and heuristics to extract candidate entity mentions (of formations, measures, among others), as well as possible co-reference relationships between the mentions.

Statistical processing. The signals produced by feature extraction may contain imprecision or inconsistency. To make coherent predictions, the developer provides constraints and (probabilistic) correlations over the signals. Hazy uses the Markov logic language; a Markov logic program consists of a set of weighted logical rules that represent high-level constraints or correlations. The developer may also specify available training data, which Hazy would use to learn rule weights. The programming interface isolates the internals of Hazy's statistical processing from the developer. It is in Hazy's infrastructure where the developer can plug in various algorithms.

Output probabilistic predictions. The output from Hazy's statistical processing infrastructure consists of probabilistic predictions on relations of interest (for example, *LocatedIn* in Figure 2). In general, Hazy prefers algorithms with theoretical guarantees (for example, Gibbs sampling). Such algorithms ensure the output predictions are well calibrated (for example, if all predictions with probability 0.7 are examined, then close to 70% of these predictions are correct). These predictions can then be fed into the front end of GeoDeepDive (Figure 3).

In addition to GeoDeepDive, we have deployed the Hazy approach in several other projects in a similar manner—for example, DeepDive ([\[hazy.cs.wisc.edu/deepdive\]\(http://hazy.cs.wisc.edu/deepdive\)\), which enhances Wikipedia with facts extracted from the Web¹⁶ \(see Figure 4\).](http://</p>
</div>
<div data-bbox=)

Programming Abstractions

Programming abstractions decouple the developer's application-specific (logical and statistical) modeling from the (statistical inference or learning) algorithms to be used for an application at execution time. The purpose of such abstractions is to ensure: an application developer can try many different algorithms for the same dataset and/or domain knowledge or heuristics with-

out additional development effort; and when the efficiency or quality of one algorithm improves, all applications using this algorithm experience automatic improvement. A combination of the relational data model and a probabilistic logic-based programming language has proved effective for meeting these two criteria.

Relational data model. As seen in the GeoDeepDive example, Hazy's approach to statistical data analysis uses the relational data model as the basis for the programming abstractions. Apart from being well studied,

Figure 3. Screen shot showing probabilistic predictions in GeoDeepDive.

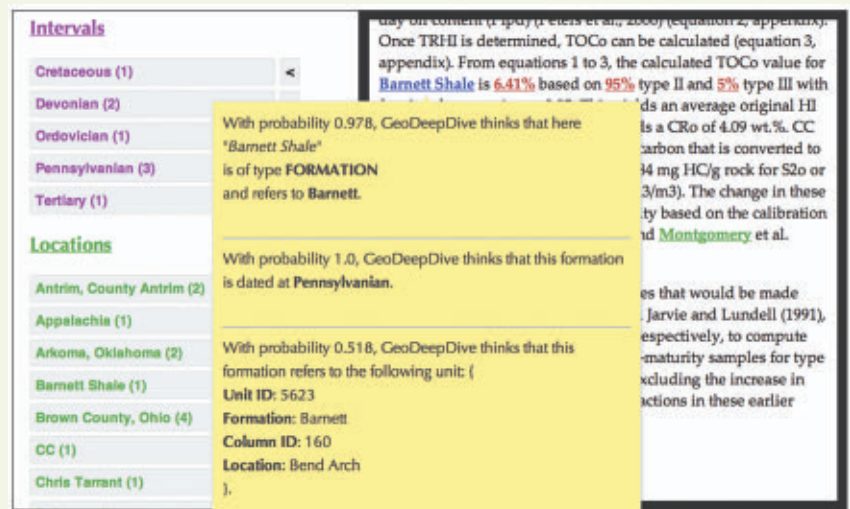


Figure 4. Sample relations about Barack Obama, Elon Musk, and Microsoft extracted by DeepDive.

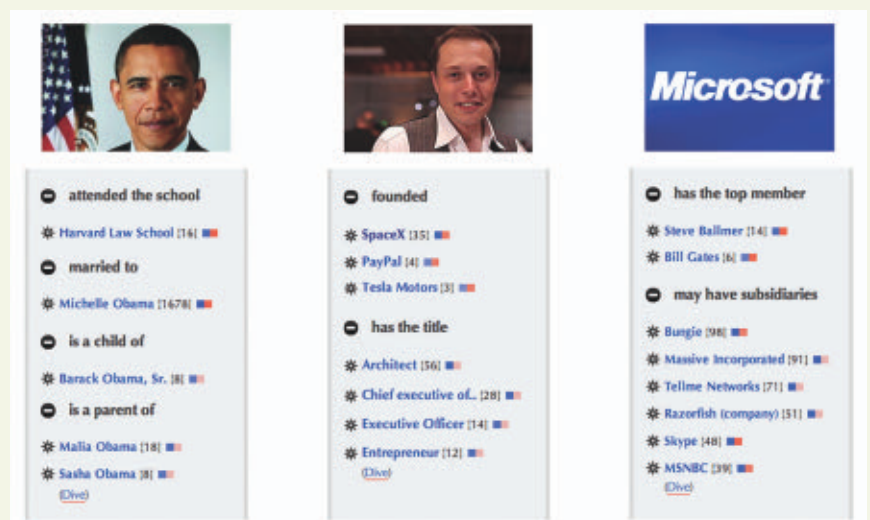
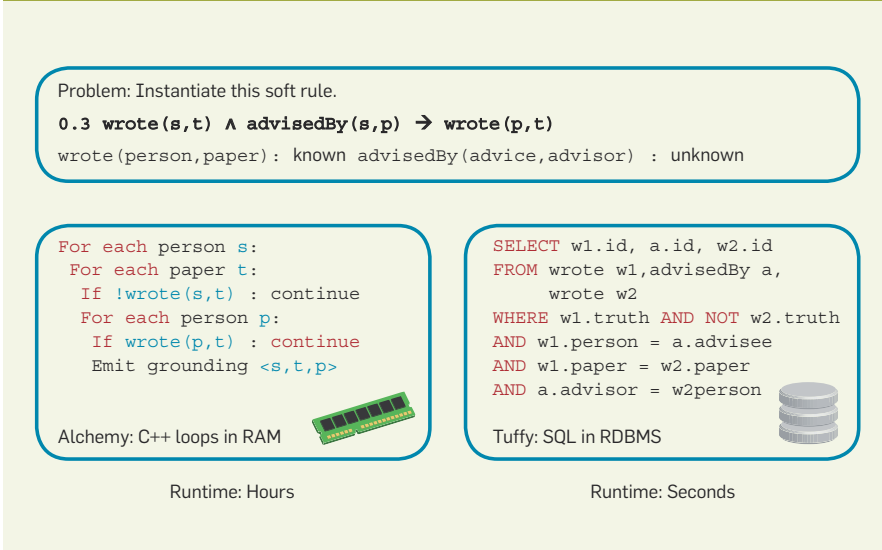


Figure 5. Comparison of in-memory grounding of Markov logic with in-RDBMS grounding.



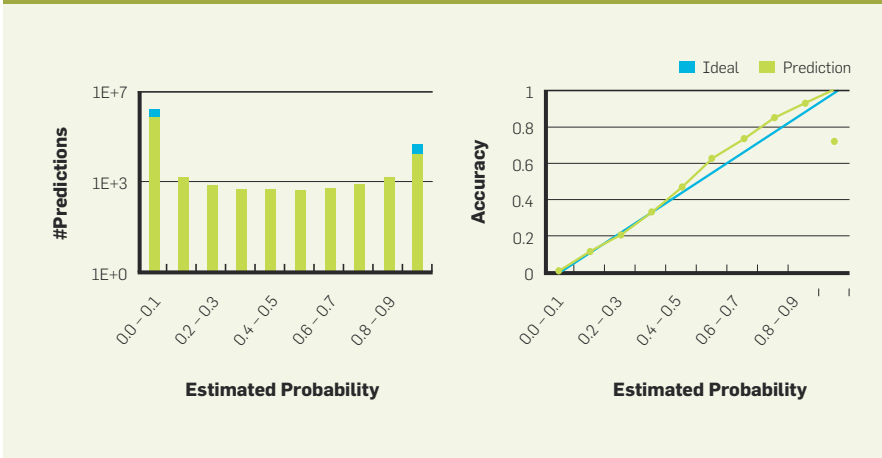
such succinctness also introduces a technical challenge of efficiently instantiating (or grounding) the first-order rules. Our crucial observation is that grounding fundamentally does relation-style joins. We built an RDBMS-based MLN interface engine, Tuffy, which leverages time-tested RDBMS infrastructure for joins to achieve high performance at scale.¹⁴ As it turned out, Tuffy was much faster and more scalable than the state-of-the-art MLN inference engine at the time (see Figure 5). The comparison is between Alchemy’s in-memory grounding (that is, rule instantiation) of Markov logic and Tuffy’s in-RDBMS grounding. Both code snippets are automatically generated by the corresponding systems for the afore-mentioned MLN rule. The use of an RDBMS makes Tuffy scalable and orders-of-magnitude faster than Alchemy, since Tuffy leverages mature RDBMS infrastructure for joins.

Markov logic is a flexible language, allowing the developer to easily represent common statistical models such as logistic regression and conditional random fields; furthermore, the developer can build more sophisticated statistical models by combining multiple “primitive” models or adding additional correlations or constraints.¹⁸

Internally, the Hazy infrastructure is able to recognize certain “primitive” models in an MLN and select inference or learning algorithms accordingly. Still, some useful statistical modeling elements are not easily represented in Markov logic (for example, correlations involving continuous random variables or aggregations). To support these more sophisticated modeling functionalities, we are extending Hazy’s programming interface to support general factor-graph construction. We are also working on extending the framework to support user-defined functions for richer application-specific logic.

Debugging. From our experience with GeoDeepDive and DeepDive, we have found debugging to be a key task in developing a trained system. Debugging is the process of performing corrections or fine-tuning to the components of an application (say, a feature extractor or an MLN program). It is error-prone and often tedious. To facilitate the debugging process, we con-

Figure 6. Macro-debugging: Example probability calibration graphs for a text-chunking task.



the relational model also underlies a large class of important statistical and machine-learning methods that use relational-style feature vectors. As an important consequence, this choice automatically provides the advantages of a mature data platform such as an RDBMS. For example, using an RDBMS to manage the data in a Hazy pipeline (as in Figure 2), a developer can easily perform data loading from and to other systems. Moreover, as RDBMS technologies continue to mature and evolve, the same Hazy pipeline would continue to gain in scalability and performance automatically.

Probabilistic logic programming. The intuitiveness, flexibility, and growing popularity of Markov logic¹⁸ led to its adoption as a central programming language in Hazy. Researchers have applied it to a wide range of applications. In Markov logic, a developer can write

first-order logic rules with weights (which intuitively model one’s confidence in a rule); this allows the developer to capture rules that are likely, but not certain, to be correct. A Markov logic program (also known as Markov logic network, or simply MLN) specifies what data (evidence) is available, what predictions to make, and what constraints and correlations exist. The process of computing predictions given an MLN is called *inference*. Sometimes an MLN may be missing weights, and a developer can provide training data from which Hazy can *learn* rule weights.

Semantically, an MLN represents a probabilistic graphical model (conceptually via rule instantiation) that in turn represents a probabilistic distribution over all possible configurations of the relations in an application. Thus, a key advantage of Markov logic is its succinctness. On the other hand,

sider it to be an integral component in programming. Here are two types of debugging that are applicable for statistical data processing in general:

► *Macro-debugging with calibration graphs.* For probabilistic predictions to make sense, they must be well calibrated. For example, if a system outputs a prediction with probability 0.7, we want the accuracy of this prediction to be 70%. A calibration graph characterizes how prediction accuracy changes with respect to prediction probabilities. In Figure 6 the x -axis is the probability of predictions estimated, and the y -axis on the left is the number of predictions made by the system (accuracy of predictions). Intuitively, if the system outputs a prediction with probability 0.7, we want the accuracy of this prediction to be 70%. These results are for a skip-chain CRF (conditional random field) model used on the CoNLL-2000 (Conference on Computational Natural Language Learning) text-chunking task, which contains a training set used for training the model, and a testing set used for evaluation. Gibbs sampling runs until convergence (decided by the Wald test) and provides inference results on the testing set. Such assessment serves as a sanity check of the whole system; if we discover that a system is not well calibrated, we can look into the training-data acquisition process and check possible overfitting problems.

► *Micro-debugging with error analysis.* To refine or add more probabilistic rules, an effective approach is to analyze errors in the results that our system produces: a developer annotates each prediction as either correct or incorrect, classifies errors into different groups, and then addresses the error groups accordingly. The challenge is that to annotate one prediction, we may need to consult many related relations. The saving grace is that, with a modeling language such as Markov logic, it is possible to trace backward from each prediction to the originating signals (and the rules in between). We are trying to design and implement a debugging IDE (integrated development environment) to support such explanations using provenance information.

Bismarck: A Unified Architecture for in-RDBMS Analytics

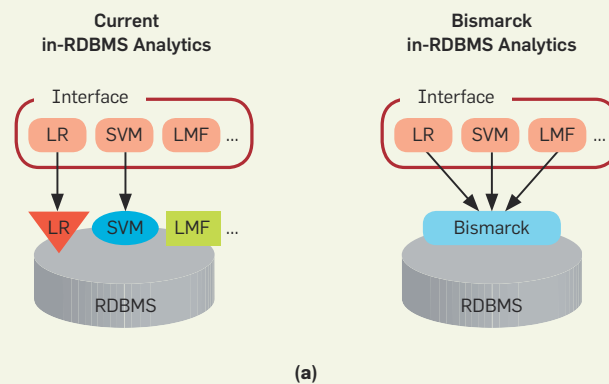
The Bismarck project¹⁰ is the first step in devising common infrastructure abstractions. Infrastructure abstractions are what decouple algorithms from implementation details such as data management, memory management, and task scheduling. Having a clean infrastructure abstraction ensures a system builder does not have to reinvent or reengineer the wheel when adding a new algorithm into the system; and when one component of the infrastructure is improved (for example, better memory management), all algorithms benefit automatically. Furthermore, a clean infrastructure abstraction provides clear angles to investigate generic techniques for improving broad classes of algorithms.

Motivation. The Bismarck project was motivated by the trend of bringing sophisticated data analytics into enterprise applications that depend on an RDBMS. From our conversations with engineers from Oracle and EMC Greenplum,¹³ we learned that

the overhead of building each new analytics technique from scratch—implementing a new solver with new memory requirements, data access methods, and others—was a major bottleneck in practice. Bismarck aims to simplify such systems with a *unified infrastructural abstraction* to handle many techniques.

Convex programming: A unifying mathematical abstraction. We begin with an important observation from the math programming literature: many analytics techniques can be framed as *convex programming problems*.^{8,12} A convex program is an optimization problem where the objective function is convex (bowl-shaped). Examples include logistic regression, SVMs (support-vector machines), and conditional random fields. Not all problems are convex (for example, Apriori¹ and some graph-mining algorithms), but a large class of problems are convex (or convex relaxations), as illustrated in Figure 7a. In contrast to existing in-RDBMS analytics tools that have separate code paths for dif-

Figure 7. (a) Bismarck in an RDBMS; (b) An incomplete list of tasks and techniques.



Analytics Task or Technique

Logistic Regression (LR)
Support Vector Machine (SVM)
Low-Rank Matrix Factorization (LMF)
Conditional Random Field (CRF)
Least-squares, Lasso, and Ridge Regression
Graph Max-Cut Problems
Kalman Filters
Portfolio Optimization

(b)

ferent analytics techniques, Bismarck provides a single framework to implement them, while possibly retaining the same interfaces. Figure 7b shows an incomplete list of tasks and techniques that can be handled by Bismarck using convex programming (and convex relaxations).

This observation is very important in data-analysis theory, since researchers are able to unify their algorithmic and theoretical studies of such problems. Convex problems are attractive since local solutions are always globally optimal, and there are many well-studied algorithms that can solve them. Because convex programming allows the problem definition to be decoupled from the way it is solved or implemented, it is a natural starting point for a unified analytics architecture.

Many analytics techniques have convex objective functions that are also linearly separable: formally, the problem is to find a vector $w \in \mathbb{R}^d$ (the *model*) for some $d \geq 1$ that minimizes the following objective:

$$\min_{w \in \mathbb{R}^d} F(w) = \sum_{i=1}^N f(w, z_i) + P(w)$$

The objective function $F(w)$ is a sum of terms $f(w, z_i)$ for $i = 1, \dots, N$ where each z is a (training) data point. In Bismarck, the z_i is represented by database tuples—for example, (paper, area) for paper classification. We abbreviate $f(w, z_i) = f_i(w)$. For example, in SVM classification, $f_i(w)$ is the hinge loss of the model w on the i th data point, and $P(w)$ enforces the smoothness of the classifier (preventing overfitting). We can generalize this to include constraints via proximal point methods. One can also generalize to both matrix valued w and nondifferentiable functions.¹⁹

Gradient methods and incremental gradient descent. There are many well-studied algorithms to solve convex programs, and the most popular are the *gradient methods*. A *gradient*, formally denoted by ∇F , is the generalization of the derivative of a function. Essentially, it gives the slope of the curve at a point, as shown in Figure 8a. The gradient is linear, which means ∇F can be computed as the sum of the N individual gradients $\nabla f(w, z_i)$.

Gradient methods are iterative algorithms that solve convex programs (1). They start at some initial value for

w and then compute the gradient (and/or related quantities) and use it to take a step to the next value of w , until the method converges to an optimum. Popular gradient methods include Conjugate Gradient, Newton Method, and BFGS.⁸ They all scan the full dataset at each iteration to compute the full gradient ∇F for a single step. This could make them inefficient for big data. Our goal is to choose a gradient method whose data-access properties are amenable to an efficient in-RDBMS implementation. A classical algorithm called IGD (*Incremental Gradient Descent*) fits the bill. IGD approximates the full gradient ∇F using only one term at a time. Formally, assuming $P = 0$ for simplicity, IGD updates the current value at iteration k , $w^{(k)}$ using a rule such as:

$$w^{(k+1)} = w^{(k)} - \alpha_k \nabla f(w^{(k)}, z_j)$$

Here, $\alpha_k \geq 0$ is a parameter called *step-size*, while z_j is one data point. In a database, each z_i corresponds to one tuple, which brings us to our central observation: IGD has a tuple-at-a-time data-access pattern that is essentially identical to a SQL aggregate such as AVG. Essentially, IGD looks at the data tuple one at a time and performs a (noncommutative) “aggregation.” IGD is also a fast algorithm, with a runtime that is linear in both the dataset size and dimension. Not surprisingly, IGD has recently become popular in the Web-data and large-scale learning communities.^{6,20}

IGD and user-defined aggregates.

The key systems insight in Bismarck is that IGD can be implemented using a classic RDBMS abstraction called a UDA (*user-defined aggregate*), which is available in almost every major RDBMS. We prototyped the same Bismarck architecture over PostgreSQL and two commercial RDBMSs. Using a UDA allows Bismarck automatically to leverage mature RDBMS capabilities such as memory management and data marshaling. It also means Bismarck can automatically leverage improvements to the RDBMS as its software evolves.

Figure 8b explains how the core computation in IGD maps to a UDA by comparing it with a SQL AVG. The state is the context of aggregation (the model in IGD). The data is a database tuple. The UDA has three stan-

Figure 8. (a) Gradient descent on a convex function; (b) Comparing AVG and IGD as a UDA.

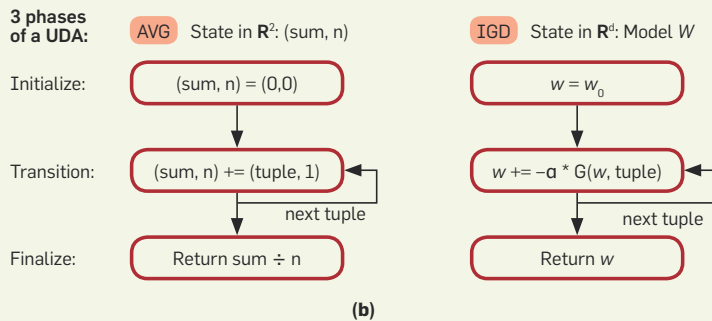
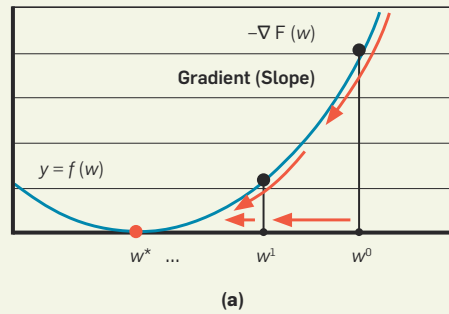


Figure 9. Snippets of C-code implementations.

```

LR_Transition(ModelCoef *w, Example e) {
    wx = Dot_Product(w, e.x);
    sig = Sigmoid(-wx * e.y);
    c = stepsize * e.y * sig;
    Scale_And_Add(w, e.x, c); ... }

```

```

SVM_Transition(ModelCoef *w, Example e) {
    wx = Dot_Product(w, e.x);
    c = stepsize * e.y;
    if(1 - wx * e.y > 0) {
        Scale_And_Add(w, e.x, c); } ... }

```

dard functions:

- `Initialize(state)` initializes the model with given values (for example, zeros, or a previous model).

- `Transition(state, data)` automatically executes on each tuple. This is where the core logic (objective function and gradient computations) of the various analytics techniques lies. Thus, the main differences between the implementations of various techniques occur primarily in a few lines of code here (Figure 9), to be explained shortly. Most of the rest of the architecture is common across techniques. This can help reduce development overhead of in-database analytics in Bismarck, in contrast to most existing systems that have a different code architecture for each technique.

- `Finalize(state)` returns the model, possibly persisting it.

A key difference of IGD from aggregates such as AVG is that IGD may need multiple passes (called *epochs*⁷) over the dataset to reach an optimum solution. The number of epochs needed is either given or determined using heuristic convergence tests based on the objective function or gradient's value.³ Another detail is that Bismarck may randomly reorder the data to improve the convergence rate of IGD.

With Bismarck's unified architecture we could rapidly implement and evaluate four popular analytics techniques—LR (logistic regression), SVM (support vector machine), LMF (low-rank matrix factorization), and CRF (conditional random field)—over three RDBMSes in less than two man-months. This is because, as mentioned earlier, a large fraction of the code infrastructure is common and reusable (on a given RDMBS). For example, starting with a full implementation of LR in Bismarck (in C, over PostgreSQL), fewer than two dozen lines of code need to change to add SVM. (The code, datasets, and a virtual machine with Bismarck preinstalled are available

for download: <http://hazy.cs.wisc.edu/victor/bismarck-download/>.) Figure 9 shows a code snippet comparison of the Transition steps of LR and SVM, where the main differences lie. Here, w is the coefficient vector, and e is a training example with feature vector x and label y . `Scale_And_Add` updates w by adding to it x multiplied by the scalar c . Note the minimal differences between the two implementations.

Similarly, more sophisticated tasks such as LMF were added with only five dozen new lines of code. This is possible since Bismarck abstracts out the invariants of the implementations of the various techniques into a small number of generic functions. This is in contrast to most existing tools, where there is usually a dedicated code stack for each technique. Apart from reducing the development overhead, the simplicity and reusability of Bismarck's architecture enables generic systems-level performance optimizations that apply to many analytics techniques. They enable Bismarck to achieve competitive (often superior) performance against many state-of-the-art commercial and open source tools on many tasks. More importantly, Bismarck achieves automatic scalability to large-scale data, as explained in the next section.

Scalability. For big-data applications, scalability is a central challenge, but Bismarck's architecture is able to achieve scalability seamlessly. Recall that Bismarck makes full use of the

powerful RDMBS abstraction of a user-defined aggregate (Figure 8b). The UDA mechanism is an industry standard that has matured over decades of development in RDBMS infrastructure. On a single node, UDAs can scale to as much data as the disk(s) can hold (hundreds of gigabytes on modern machines), but UDAs are also scalable to a parallel database cluster with one addition to the three-function abstraction of Figure 8b: a `Merge(state, state)` function that merges partial aggregates computed on partitioned data in shared-nothing nodes. Although IGD is not algebraic¹¹ as SQL AVG, we can leverage model-averaging ideas from the literature.²¹ Thus, Bismarck offers automatic scalability for many analytics techniques in its unified architecture. In contrast, in many custom-built systems, scalability is either not taken into consideration at all, or the developers will have to worry about implementing data management and scalability issues, often reinventing the wheel.

The accompanying table shows some experimental results validating Bismarck's scalability. (A more comprehensive experimental evaluation is available in the Bismarck paper.¹⁰) A check mark means the task completes, and X means the approach either crashes or takes longer than 48 hours. N/A means the task is not supported. We compare Bismarck over PostgreSQL against the native analytics tool of a commercial engine DBMS A, as well as

Bismarck scalability.

Task	Dataset			Bismarck PostgreSQL	DBMS A (Native)	In-memory Tools
	Name	#Examples	Size			
LR	Classify300M	300M	135GB	✓	✓	X
SVM				✓	✓	X
LMF	Matrix5B	5B	190GB	✓	N/A	X
CRF	DBLP	2.3M	7.2GB	✓	N/A	X

popular task-specific in-memory tools (Weka, SVMPerf, CRF++, Mallet). All in-memory tools crashed either due to insufficient memory (Weka, SVMPerf, CRF++) or did not terminate even after 48 hours due to thrashing (Mallet). All of the in-RDBMS tools can scale on the simple tasks LR and SVM (less than an hour per epoch for Bismarck), and Bismarck also scales on more complex tasks that are not currently available in DBMS A.

HogWild! and Jellyfish: Extending the infrastructure abstraction. An infrastructure abstraction such as Bismarck's enables us to study generic system optimizations that apply to many techniques. One such optimization is the HogWild! mechanism to parallelize IGD.¹⁵ Modern machines (say, in enterprise settings) typically have multiple cores and shared memory accessible by each core. IGD can then run in parallel on these cores, with the model residing in shared memory.

While one might think locking is required to avoid race conditions, the HogWild! approach is not to lock at all. Under some assumptions, HogWild! guarantees convergence and similar quality. Such lockfree parallelism means HogWild! can achieve near-linear speedups on all the analytics techniques in Figure 7b. We also integrated the HogWild! idea into Bismarck as an alternative to the native UDA parallelism (shared-nothing) and found the former to be significantly faster for large data that can reside on a single node.¹⁰

Some analytics techniques have specific structures, which can be further exploited to improve performance. Matrix factorization is an excellent example, where the Jellyfish mechanism exploits the structure and outperforms HogWild!. Jellyfish achieves this speedup by using the Latin square pattern to chunk the data matrix.¹⁷ Unlike HogWild!, such chunking enables Jellyfish to run the factorization in parallel on multiple cores without any concurrent overwrites or averaging needed on the model.

Overall, as our experiences with Bismarck, HogWild!, and Jellyfish (and its successor HotTopixx⁵) show, fundamental infrastructure abstractions simplify the development of trained systems by decoupling the algorithms



For big-data applications, scalability is a central challenge, but Bismarck's architecture is able to achieve scalability seamlessly.



from their implementations. Such decoupling allows us to leverage existing mature code infrastructures such as UDAs, automatically providing features such as maintainability and scalability. It also allows us to devise new performance optimizations that can be designed once and applied to many techniques, rather than reinventing the wheel again and again. While we have focused on applications that use an RDBMS here, the infrastructure abstraction offered by Bismarck is also amenable to newer data platforms such as MapReduce/Hadoop that offer UDA-like aggregation capabilities. We are working on applying our lessons from Bismarck to some of these data platforms as well.

Future Work and Open Challenges

The Hazy abstractions are a continuously evolving and growing collection, and the primary source of motivation and inspiration is our own experience in developing and deploying trained systems such as GeoDeepDive and DeepDive. As we refine existing abstractions and explore new ones, the following challenges may be particularly interesting (and we are actively working in these directions):

Feature engineering. Conventional wisdom goes that “more signals beat sophisticated models,” and our experience in developing GeoDeepDive affirms this idea. Thus, features (or statistical signals) could have a first-class-citizen status just as algorithms in a framework such as Hazy. In contrast to algorithms that are typically off-the-shelf, the effectiveness of features are usually application dependent. As a result, the process of feature engineering tends to be iterative and have humans in the loop. We are trying to abstract feature engineering as a cyclic process involving E3: (data) exploration, (feature) extraction, and (results) evaluation.

Assisted development. Traditionally, developing trained systems requires expertise, experience, and a deep understanding of the data and algorithms. As a result, usually only a small number of developers would feel qualified for such applications; and the development process would often be tricky or painstaking even for these developers. To lower the barrier and

improve the productivity of developing such applications, we are exploring various options to support assisted development (for example, automatic feature suggestion, automatic parameter tuning, and smart diagnosis of trained systems²).

New data platforms. Some recent projects aim to bring statistical tools to the Hadoop environment.^{4,9} It would be interesting to port the Hazy abstractions to Hadoop (and associated systems such as HBase and Accumulo). The resulting combination is likely to enjoy a large and active user base of developers. A key challenge in this direction is how to reconcile the roles played by the RDBMS in Hazy in the Hadoop environment. We are exploring possible solutions to address this challenge.

Conclusion

There is a race to unleash the full potential of big data using statistical and machine-learning techniques. The high-profile success of many recent big-data analytics-driven systems, also known as trained systems, has generated great interest in bringing such technological capabilities to a wider variety of domains. A key challenge in converting this potential to reality is making these trained systems easier to build and maintain.

The Hazy project outlined an approach to tackling this challenge by identifying common patterns, also known as abstractions, in building such systems. The abstractions allow for decoupling the concerns of applications from the algorithms that are used and the underlying implementations that are needed. Optimizing and supporting these abstractions as primitives enables us to make it easier to build and maintain trained systems. Our ideas are shaped by, and continue to evolve with, our own experiences with building such systems (DeepDive, GeoDeepDive, and AncientText), as well as our repeated interactions with practitioners at major enterprises and developers at major analytics companies.

The Hazy Research Group's philosophy is to make all our research software available as open source. The source code, installation and usage documentation, datasets used for our research publications, and virtual

machines with the software tools and datasets preinstalled are available at <http://hazy.cs.wisc.edu>. The tools have already been downloaded thousands of times. Videos providing overviews of Hazy projects and tutorials are available at <http://www.youtube.com/user/HazyResearch/videos>. Hazy code has been adopted and shipped into production by four companies, is used by many other research groups, and has been deployed at a scientific observatory at the South Pole.

Acknowledgments

The Hazy Research Group is a team of Ph.D., M.S., and undergraduate students, working under the supervision of Christopher Ré. The Hazy project is made possible due to the endless enthusiasm and tireless efforts of these students. In addition to the authors, the following students also contributed to this article: Victor Bittorf, Xixuan Feng, and Ce Zhang. We gratefully acknowledge the support of DARPA grant FA8750-09-C-0181, NSF CAREER award IIS1054009, ONR award N000141210041, and gifts or research awards from American Family Insurance, Google, Greenplum, Johnson Controls, Inc., LogicBlox, Oracle, and Raytheon. We appreciate the support of the Center for High Throughput Computing and Miron Livny's Condor research group. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the above companies, DARPA, or the U.S. government. ■

Related articles on queue.acm.org

The Pathologies of Big Data

Adam Jacobs

<http://queue.acm.org/detail.cfm?id=1563874>

How Will Astronomy Archives Survive the Data Tsunami?

G. Bruce Berriman, Steven L. Groom

<http://queue.acm.org/detail.cfm?id=2047483>

Condos and Clouds

Pat Helland

<http://queue.acm.org/detail.cfm?id=2398392>

References

1. Agrawal, R. and Srikant, R. Fast algorithms for mining association rules in large databases. In *Proceedings of Very Large Databases*, 1994.
2. Anderson, M., Antenucci, D., Bittorf, V., Burgess, M., Cafarella, M., Kumar, A., Niu, F., Park, Y., Ré, C. and Zhang, C. 2013. Brainwash: A data system for

feature engineering. In *Proceedings of Conference on Innovative Data Systems Research*, 2013.

3. Anstreicher, K.M., Wolsey, L.A. Two "well-known" properties of subgradient optimization. *Mathematical Programming* 120, 1 (2009), 213–220.
4. Apache Mahout; <http://mahout.apache.org/>.
5. Bittorf, V., Recht, B., Ré, C. and Tropp, J. Factoring nonnegative matrices with linear programs. In *Proceedings of Neural Information Processing Systems*, 2012.
6. Bottou, L. and Bousquet, O. The tradeoffs of large scale learning. In *Proceedings of Neural Information Processing Systems*, 2007.
7. Bottou, L. and LeCun, Y. Large scale online learning. In *Proceedings of Neural Information Processing Systems*, 2003.
8. Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, NY, 2004.
9. Das, S., Sismanis, Y., Beyer, K. S., Gemulla, R., Haas, P. J. and McPherson, J. Ricardo: Integrating R and Hadoop. In *Proceedings of ACM SIGMOD*, 2010.
10. Feng, X., Kumar, A., Recht, B. and Ré, C. Towards a unified architecture for in-RDBMS analytics. In *Proceedings of ACM SIGMOD*, 2012.
11. Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., and Pirahesh, H. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data Mining and Knowledge Discovery* 1, 1 (1997).
12. Hastie, T., Tibshirani, R. and Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, NY, 2011.
13. Hellerstein, J., Ré, C., Schoppmann, F., Wang, D. Z., Fratkín, E., Gorajek, A., Ng, K. S., Welton, C., Feng, X., Li, K. and Kumar, A. The MADlib Analytics Library or MAD Skills, the SQL. In *Proceedings of the VLDB Endowment* 5, 12 (2012): 1700–1711.
14. Niu, F., Ré, C., Doan, A. and Shavlik, J. Tuffy: Scaling up statistical inference in Markov logic networks using an RDBMS. In *Proceedings of Very Large Databases*, 2011.
15. Niu, F., Recht, B., Ré, C. and Wright, S. Hogwild!: a lock-free approach to parallelizing stochastic gradient descent. In *Proceedings of Neural Information Processing Systems*, 2011.
16. Niu, F., Zhang, C., Ré, C. and Shavlik, J. Elementary: Large-scale knowledge-base construction via machine learning and statistical inference. *International Journal on Semantic Web and Information Systems-Workshop on Web-scale Knowledge Extraction*, 2012.
17. Recht, B. and Ré, C. Parallel stochastic gradient algorithms for large-scale matrix completion. In *Optimization Online*, 2012.
18. Richardson, M. and Domingos, P. Markov logic networks. *Machine Learning* 62 (2006), 107–136.
19. Rockafellar, R.T. *Convex Analysis* (Princeton Landmarks in Mathematics and Physics). Princeton University Press, Princeton, NJ, 1996.
20. Vowpal Wabbit; <http://hunch.net/~vw/>.
21. Zinkevich, M., Weimer, M., Smola, A. and Li, L. Parallelized stochastic gradient descent. In *Proceedings of Neural Information Processing Systems*, 2010.

Arun Kumar is a Ph.D. student at the University of Wisconsin-Madison. His research interests are in the areas of data management, with a focus on data analytics and managing uncertain data. He received his M.S. from the University of Wisconsin-Madison in 2011, and his B.Tech from the Indian Institute of Technology-Madras in 2009, both in computer science.

Feng Niu is a software engineer at Google, Inc. His goal is to help the machine help organize the world's information with more structures, connections, and insights, but with less human effort. In 2012, he received his Ph.D. in computer science from the University of Wisconsin-Madison. His graduate study was mainly funded by the DARPA Machine Reading program. He received his BE degree in Computer Science from Tsinghua University in 2008.

Christopher (Chris) Ré is an assistant professor in the department of computer sciences at the University of Wisconsin-Madison. The goal of his work is to enable users and developers to build applications that more deeply understand and exploit data. He received his Ph.D. from the University of Washington, Seattle; his work in the area of probabilistic data management received the SIGMOD 2010 Jim Gray Dissertation Award. Ré received an NSF CAREER Award in 2011.

© 2013 ACM 0001-0782/13/03

It is easy to do amazing things, such as rendering the classic teapot in HTML and CSS.

BY BRIAN BECKMAN AND ERIK MEIJER

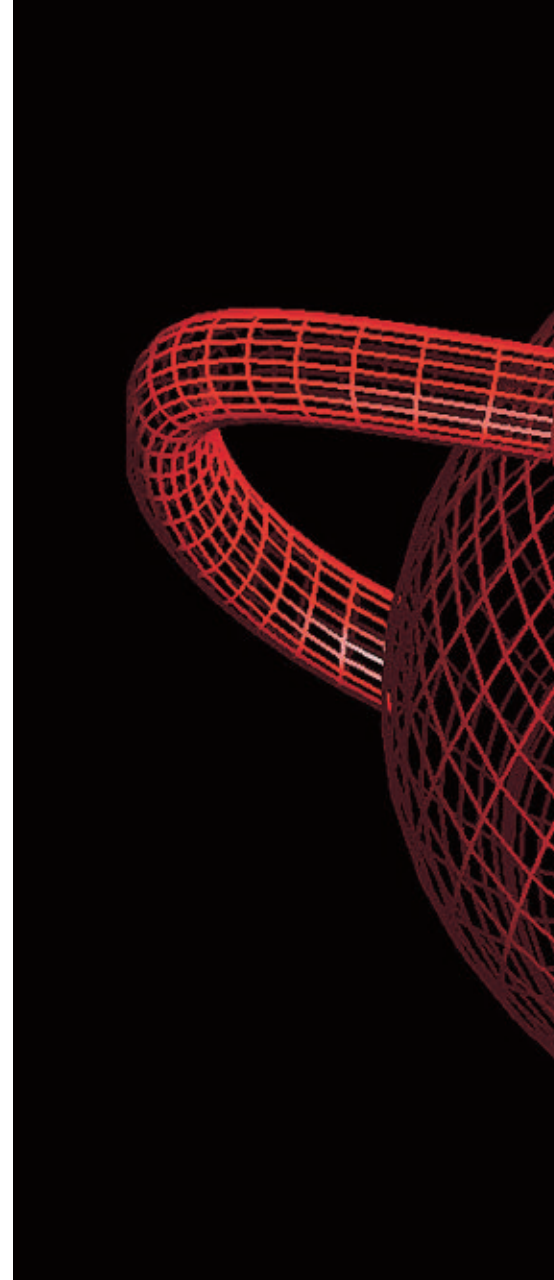
The Story of the Teapot in DHTML

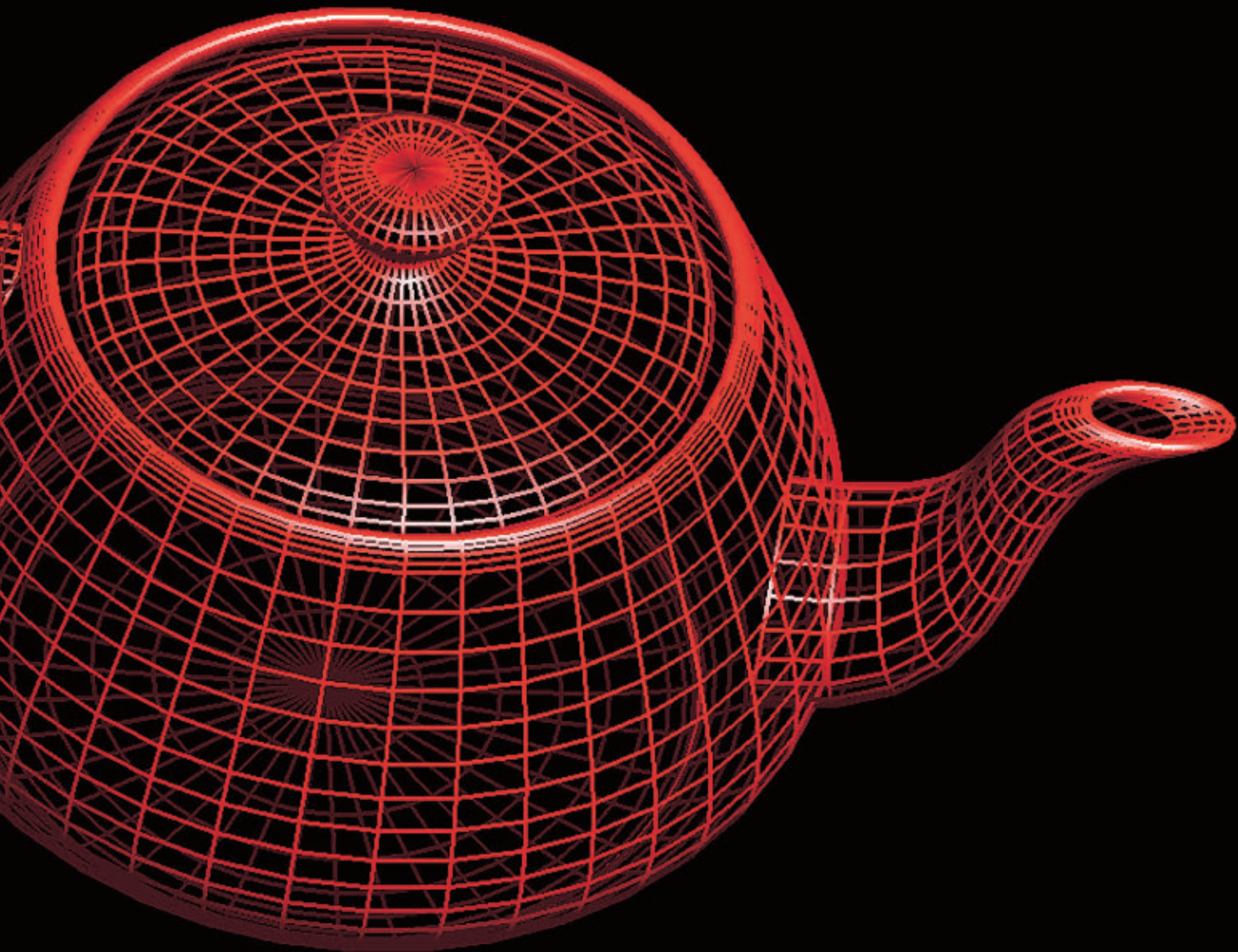
BEFORE THERE WAS Scalable Vector Graphics (SVG), Web Graphics Library (WebGL), Canvas, or much of anything for graphics in the browser, it was possible to do quite a lot more than was initially obvious. To demonstrate, we created a JavaScript program that renders polygonal 3D graphics using nothing more than HTML and CSS. Our proof-of-concept is fast enough to support physics-based small-game content, but we started with the iconic 3D “Utah teapot” (Figure 1) as it tells the whole story in one picture. (For background, see <http://bit.ly/KQK9a>.) It is feasible to render this classic object using just regular `<div>` elements, CSS styles, and a bit of JavaScript code (Figure 2). This tiny graphics pipeline serves as a timeless demonstration of doing a lot with very little.

The inspiration for this project came from Web developer Jeff Lau, who on his blog UselessPickles implemented a textbook graphics pipeline in handwritten JavaScript (<http://www.uselesspickles.com/triangles/>)

Lau’s demo embodies a glorious hack for efficiently rendering triangles in HTML, shown in Figure 3.

The glorious hack is explained later, but, to spoil the punch line, once you have arbitrary triangles, you can easily render arbitrary polygons, and thus arbitrary polygon-based models. The only remaining issues for game-competent 3D graphics are texture mapping, bump mapping, reflection mapping, and performance. These various kinds of mappings all require pixel-based primitives: the ability to render individual pixels efficiently. Though it is *possible* to render individual pixels using just the `<div>` element with a CSS style to shrink the element to pixel size, this obviously does not provide sufficient performance for classic scan conversion of 3D models. The work to render individual pixels is quadratic in





the linear size of a 2D figure, meaning that doubling the size of a figure requires roughly four times the work.

The `<div>` elements, however, also begrudgingly provide a way to draw vertical and horizontal lines, if, for no other reason, than for borders around text. Several other bloggers (including David Betz and the late Walter Zorn) noted that by decomposing a figure into parallel “raster lines” instead of pixels, work is linear in the linear size of the figure, meaning that doubling the size only doubles the work. They created JavaScript 2D graphics libraries with reasonable performance by combining *pixel drawing where necessary and line drawing where possible* in `<div>`s.

The following is an HTML page that illustrates the linear method by drawing a right triangle of eight vertical raster lines of linearly increasing height:

```
<style>div{ background:Black;
position:absolute; width:9px; }
</style>
<div style="left:10px;
height:10px;"></div>
<div style="left:20px;
height:20px;"></div>
<div style="left:30px;
height:30px;"></div>
...
<div style="left:80px;
height:80px;"></div>
```

The CSS style sheet and the inline position and height declarations create eight instances of `<div>` with linearly increasing left coordinate and linearly increasing height. This HTML page renders as shown in Figure 4.

The linear pattern of coordinate and height values in HTML should be obvious. It should also be obvious how

to write a program to generate a similar HTML page that renders not only right triangles with a scheme like this: just arrange the coordinate and dimension values to be linearly increasing or decreasing in appropriate ways. The following program dynamically generates `<div>` elements exactly as the static markup:

```
<script>
  for(var i = 1; i < 9;
  i++)
    with(document)
      with(body.appendChild(
        createElement("div")).
        style)
        {
          left = i * 10;
          height = i * 10;
        }
</script>
```

Figure 1. The “Utah Teapot” (from “Fast Ray Tracing by Ray Classification,” by James Arvo and David Kirk, 1987).



Figure 2. The teapot rendered in HTML, CSS, and JavaScript.

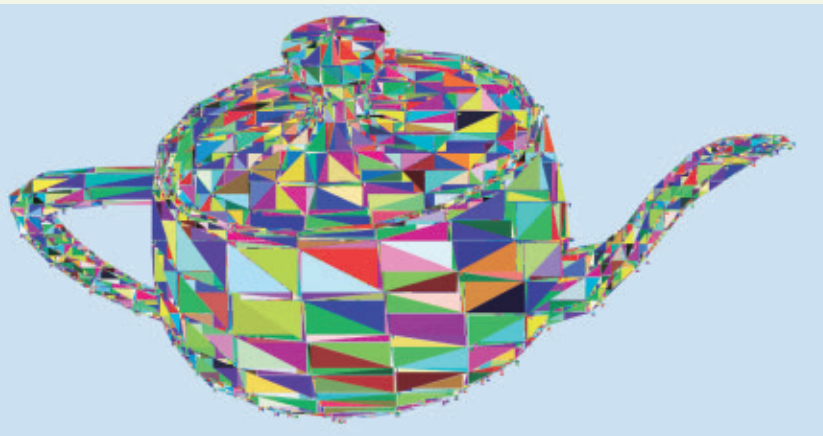
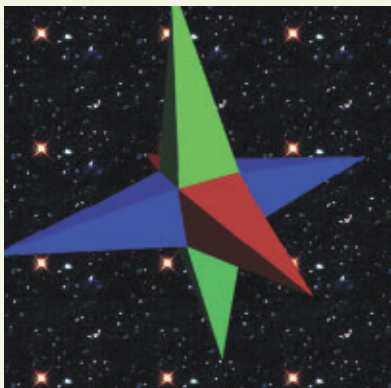


Figure 3. Lau’s DHTML demo.



Standard rasterizer algorithms such as Bresenham’s permit drawing all kinds of figures. Any programmer who has taken Computer Graphics 101 has seen enough now to create an entire workable 2D graphics library in HTML.

Logarithmic Performance

It is possible to achieve quadratic performance from pixels and linear performance from rasters. Can we do better than linear? Lau found *logarithmic* performance, meaning that doubling the linear size of a figure requires only a constant amount of more work, usually just one or two more calls to primitives. Logarithmic is *much* more efficient than linear. The difference is the same as that between binary search and linear search. Lau noticed that `<div>` + CSS has a subtly hidden primitive right triangle, if you know where to look. Then he presented a beautiful way to decompose an arbitrary triangle into a logarithmic number of right triangles: his glorious hack.

Notice in Figure 5 that HTML allows rendering the four borders of a `<div>` completely independently by setting the `border-XXX` colors. Setting the width of the `<div>` to zero removes the text, leaving just four triangles, as in Figure 6.

Is it possible to get rid of two of the triangles? This is straightforward: make the width of the right (yellow) border zero as in the “animation” in Figure 7, and similarly shrink the bottom (blue) border to make it disappear as well, as in Figure 8; this leaves just a green and a red right triangle.

Figure 4. Render of linear method.



Figure 5. HTML border colors.

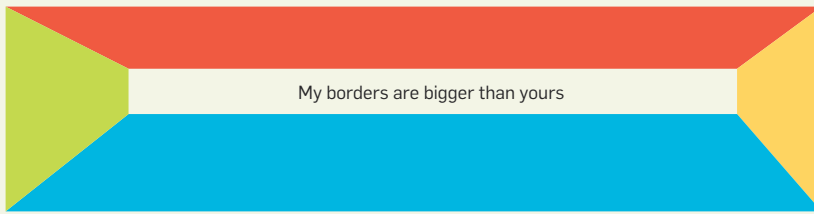


Figure 6. Triangles in HTML.



Figure 7. Shrinking the first triangle.



Figure 8. Shrinking the second triangle.



Now, setting one of the remaining border colors to “transparent” renders a single right triangle at the native efficiency of the underlying browser’s rendering engine, presumably very high (Figure 9).

Setting the left and bottom borders’ width to zero and making the other appropriate borders transparent—straightforward extensions of Lau’s method—produces HTML primitives for all four kinds of right triangles, as in Figure 10.

Assume at this point a JavaScript function `drawRightTriangle(P1, P2, P3)` that can render any of these right triangles given the (coordinates of the) three vertices. The details are tedious and unenlightening, but suffice it to say that each of the four branches in the implementation must set the proper attributes of an underlying `<div>` tag, either directly or through CSS style classes.

Now we have really fast *right* tri-

Figure 9. Making the third triangle transparent.



(a)



(b)

Figure 10. HTML primitives for all four kinds of right triangles.



(a)



(b)

angles, but where are the *arbitrary* triangles with logarithmic performance, where doubling the triangle size means just one or two extra calls to the right-triangle primitive? Lau’s original code is iterative in style, but

the underlying recursive description is elegant, as follows:

Consider an arbitrary triangle; by definition of a triangle, the three vertices are *not* all on the same line. There are just two cases to consider: either

there is one horizontal leg, or there is not (Figure 11). If there is one horizontal leg, then skip to the next paragraph. If there is not one horizontal leg, then cut the triangle with one horizontal line into *two* triangles, each with one horizontal leg. Cutting the triangle means computing the coordinates of a new point, P4, as in Figure 12.

The y coordinate of the new point P4 is the same as the y coordinate of the middle-in-y point, P2—that is, $P4.y == P2.y$ —and the x coordinate of the new point is proportionately as far from the x coordinate of the bottom point as the y coordinate of the new point is from the y coordinate of the bottom point:

$$P4.x == P3.x + (P1.x - P3.x) * ((P4.y - P3.y) / (P1.y - P3.y))$$

The pseudocode, assuming that P1, P2, and P3 are in downward, increasing-y order, is:

```
function
drawTriangleWithoutHorizontalLeg(P1, P2, P3)
{
  ... compute P4 according
  to equations above ... ;
  drawTriangleWithOneHorizontalLeg(P1, P2, P4);
  drawTriangleWithOneHorizontalLeg(P3, P2, P4);
}
```

There is one final function to write:

`drawTriangleWithOneHorizontalLeg`. Recall that the two kinds of triangles-with-one-horizontal-leg are hanging-down and standing-up. They are completely symmetric, so let us work out the final steps only for the standing-up triangle. There are three possible cases:

- ▶ The tip is between the two base vertices—an *acute* triangle.
- ▶ The tip is exactly over one of the two base vertices—a *right* triangle.
- ▶ The tip is either to the right or the left of the base segment—an *obtuse* triangle.

If *acute*, cut the triangle vertically into two right triangles and call it a day! If *right*, well, it is a right triangle and done! If *obtuse*, then cut the triangle vertically

Figure 11. Two types of triangle.

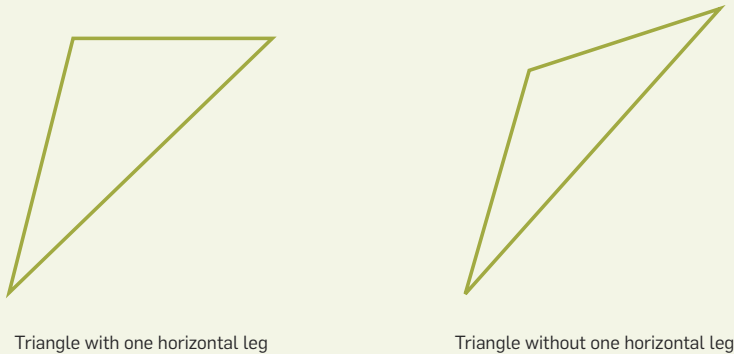


Figure 12. Forcing a horizontal side of the triangle.

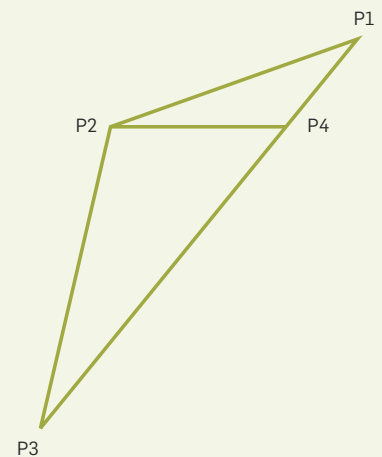


Figure 13. Forcing a vertical side of the triangle.

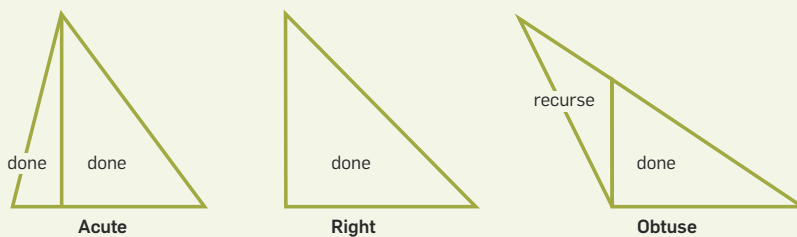


Figure 14. The decomposition of a triangle.

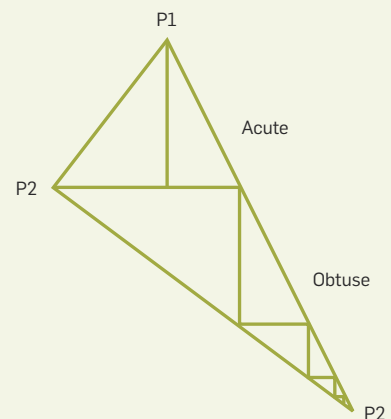
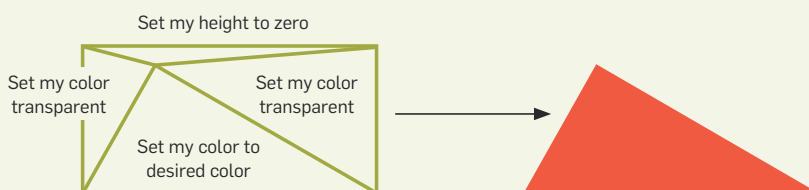


Figure 15. A small improvement.



into one right triangle (done) and one obtuse triangle (recurse on drawTriangle). Beautiful! (See Figure 13.)

To avoid infinite recursion in the third case, you must also stop if a triangle is too small—say, smaller than one pixel. From this description, all the corner cases are covered and any programmer should be able to write a correct implementation that performs well. When all the recursion has bottomed out, the decomposition of a triangle looks like Figure 14, which is what Lau drew in the first place.

His algorithm decomposes an arbitrary triangle into one or two triangles with horizontal legs; let's call those *aligned triangles*. It decomposes any acute aligned subtriangle into two aligned right triangles, and it decomposes any aligned obtuse subtriangle into a recursive number of aligned right triangles. Mathematically, this recursive number is infinite. Computationally, because you can render only a finite approximation of the mathematical structure on a screen with a finite pixel size, the number is logarithmic in the size of the figure.

Note the following small improvement: an acute aligned triangle can be rendered directly by setting the border opposite the aligned leg of the `<div>` to zero width and the borders on either side of the target triangle to transparent color, as in Figure 15.

Also note that it seems impossible to decompose an obtuse aligned triangle into a finite number of acute aligned triangles. Marc Levy provides the following argument sketch: if there were a finite number, then you could find the smallest one by sorting them by size. Consider the smallest one and draw a horizontal cut from its vertex farthest from the base of the original triangle. The residue is a triangle similar to the original triangle, thus leaving a smaller version of the original problem, in which there are even smaller acute aligned triangles. We assumed, however, that we had the smallest one, so that premise must be wrong: there does not exist a smallest one; therefore, there is not a finite number of them.

From a Teapot to Triangles and Back

Now that you know how to draw any

Table 1. Original teapot data; 3,751 triangles too much for good dynamic performance.

1976	3751	11253
-3.0000	1.65	0
-2.98711	1.65	-0.098438
-2.98538	1.56732	-0.049219
... 1976 vertex-coordinate triplets ...		
3 1455	1469	1459
3 1449	1455	1459
3 1462	1449	1459
... 3751 triangle patches as indices into the array of vertex-coordinate triplets ...		

Table 2. Teapot data after uniform decimation; 263 triangles with great performance.

166	263	0
0.000000	0.000000	0.488037
0.941342	-0.188153	0.626546
-0.000023	-0.032926	0.473905
... 166 vertex-coordinate triplets ...		
3 57	50	130
3 57	130	123
3 50	81	47
... 263 triangle patches as indices into the array of vertex-coordinate triplets ...		

triangle anywhere on the screen, how can you get the teapot? First, get some data from the public domain in a well-documented format called OFF (http://segeval.cs.princeton.edu/public/off_format.html) as seen in Table 1.

Use a free graphics tool to decimate the data (trim it down to manageable size) to get the data in Table 2. Then convert the data into JavaScript via editor macros or a simple script. The final ingredients to the graphics pipeline are backface culling, Z-ordering, orientation by quaternions, and a kinematics library for spinning the object. Add a little spring-and-dashpot physics, and you can make your teapot bounce, morph, shatter, and unshatter. Code for all of these examples is available at <https://github.com/gousiosg/teapots>.

With the right leverage applied at exactly the right fulcrum point, it is relatively easy to do amazing things, such as rendering the classic teapot in HTML and CSS. Or, to paraphrase the Hacker's dictionary, we can extract great pleasure out of stretching the capabilities of programmable systems beyond the original intent of their designers.

Related articles on queue.acm.org

Scripting Web Services Prototypes

Christopher Vincent

<http://queue.acm.org/detail.cfm?id=640158>

A Conversation with Ray Ozzie

<http://queue.acm.org/detail.cfm?id=1105674>

Mobile Application Development: Web vs. Native

Andre Charland, Brian LeRoux

<http://queue.acm.org/detail.cfm?id=1968203>

Brian Beckman (bbeckman@exchange.microsoft.com) is working with Bing on Maps and Signals and has held many positions at Microsoft since 1992, from Crypto (SET) to Biztalk to research in functional programming. He wrote the first version of the Time Warp Operating System on the Caltech Hypercube 1984–1989. He holds a Ph.D. in Astrophysics from Princeton University (1982) and has filed over 80 patents with 30 issued.

Erik Meijer (emeijer@microsoft.com) has been working on "democratizing the cloud" for the past 15 years. He is perhaps best known for his work on the Haskell, C#, and Visual Basic languages, targeting JavaScript as assembly language, and his contributions to LINQ and Rx (Reactive Framework). He is a part-time professor of cloud programming at TUDelft and runs the cloud programmability team at Microsoft.

Mobile performance issues?
Fix the back end, not just the client.

BY KATE MATSUDAIRA

Making the Mobile Web Faster

MOBILE CLIENTS HAVE been on the rise and will only continue to grow. This means that if you are serving clients over the Internet, then you cannot ignore the customer experience on a mobile device.

There are many informative articles on mobile performance, and just as many on general API design, but you will find few discussing the design considerations needed to optimize the back-end systems for mobile clients. Whether you have an app, mobile website, or both, it is likely these clients are consuming APIs from your back-end systems. It is this part of that infrastructure that this article is about.

Certainly, optimizing the on-mobile performance of the application is critical, but software engineers can do a lot to ensure mobile clients are remotely served both data and application resources in a reliably performant manner.

What is so special about mobile? If you were to go back in time and use the Internet, you would notice

that most websites felt slower. The technology has now evolved to the point that clients can efficiently use and negotiate low-bandwidth channels. Mobile clients, however, do not have the computer power, storage, and high-bandwidth connections of desktops, so mobile needs to be thought about a little differently.

Here are some of the special considerations to take into account when building mobile-based applications:

- ▶ *Limited screen size.* There is less space for data and images.

- ▶ *Smaller number of simultaneous connections.* This one is important because unlike Web browsers that can run many concurrent asynchronous requests, mobile browsers have a limited number of connections per domain at any given moment.

- ▶ *Slower network.* Network performance is heavily affected by poor signal reception and multiple cellular handovers (even though some clients are on Wi-Fi, some networks are congested and can require additional look-ups if a user changes cell towers).

- ▶ *Slower processing power.* Extensive client-side computations, 3D graphics rendering, and heavy JavaScript usage can greatly affect performance.

- ▶ *Smaller caches.* Mobile clients are generally memory-restricted so it is best not to rely heavily on cached content for performance.

- ▶ *“Special” browsers.* In many ways the mobile browser ecosystem is reminiscent of the fragmented desktop browser scene of several years ago, with mobile vendors producing versions with fatal deficiencies and incompatibilities.

Although there are many ways to tackle these unique obstacles, this article focuses on what can be done from an API or back-end service to improve the performance (or the perception thereof) of mobile clients. The article is divided into two parts:

- ▶ Minimizing network connections and the need to transmit data—efficient media handling, effective caching, and employing longer data-oriented operations with fewer connections.

► Sending the “right” data across the network—designing APIs to return only the data that is needed/requested, and optimizing for the various types of forms of mobile devices.

Although this article is focused solely on mobile, many of the lessons and ideas can be applied to other API client forms as well.

Minimizing connections and data across the network. Minimizing the number of HTTP requests required to render a Web page is undoubtedly one of the biggest ways of improving mobile performance. There are many ways to do this, but the exact approach may depend on your data and the architecture of your application.

In most cases you want to minimize how much information is sent across the network. Rendering on the server has its advantages (such as when the server sends back whole HTML pages) since it requires less compute and processing resources than doing so on the client. Of course, the downside of this approach is that the more code rendered server side, the more likely that code may have display issues in client browsers (and dealing with browser compatibility is seldom fun). Still, the more that can be done on the client, the fewer trips across the network. After all, that is why “apps” have become so popular—if you could do everything in the Web browser with the network, this would be a mobile website world.

Minimize image requests. In a standard browser, making a single request for each image on the page results in speed improvements and allows you to take advantage of caching for each image. The browser is able to execute each request quickly and in parallel, so there is not a big performance hit for making many requests (and with the caching benefits there can even be performance gains). This same request, however, can be a killer on mobile.

Since every request for data on a mobile device can require substantially more overhead, it can add significant latency to each request. There-

fore, minimizing image requests can reduce the number of requests and in some cases the amount of data that needs to be sent (which can also help mobile performance).

Here are some strategies to consider:

Use image sprites. The use of image sprites can reduce the number of individual images that need to be downloaded from the server, but sprites can be cumbersome to maintain and difficult to generate in some circumstances (such as on product search results where you are showing thumbnail images for many products).

Use CSS instead of images. Avoiding images where possible and using CSS (Cascading Style Sheets) rendering for shadows, gradients, and other effects can reduce the amount of bytes that need to be transmitted and downloaded.

Support responsive images. A popular way of delivering the right image to the right device is using responsive images. Apple does this by loading regular images and then replacing them with high-resolution ones using

JavaScript.⁷ There are several other ways³ of approaching this problem, but the issue is far from solved.¹²

In these cases you should make sure that the server-side support and APIs are able to support different versions of the same image, and the exact way to do that will depend on the approach of the clients. For example, one easy way of doing this with an API is to support a handful of image sizes as a parameter for the request, as shown in Figure 1.

To keep APIs simple, make this parameter optional and send back a default size. To pick your default size, select either the smallest size (to handle situations such as responsive images) or the most commonly used size on your website.

Use data URIs for images inline to minimize extra requests. An alternative to sprites is to use data URIs (uniform resource identifiers) to embed images inline within the HTML itself. This makes the images part of the overall page, and while the URI-encoded images can be larger in terms of bytes, they compress better with gzip com-



Figure 1. Example request and response using a parameter to indicate image size.**Request:**

```
http://yourdomain.com/api/objects.json?objectIds=18369542&imageSize=IMG_140x140
```

Response:

```
objects: [
  {
    product: {
      id: "18369542",
      title: "Upright Freezer",
      brand: "Frigidaire",
      imageURL: "https://yourdomain.net/140x140/18369542-140x140.jpg",
    }
  },
  {
    product: {
      id: "14958145",
      title: "Sony Bravia 32" LCD",
      brand: "Sony",
      imageURL: "https://yourdomain.net/140x140/14958145-140x140.jpg",
    }
  }
]
```

These are the size options used in my last project

```
{ 'IMG_ORIGINAL' | 'IMG_70x70' | 'IMG_80x80' | 'IMG_85x85' | 'IMG_90x90' | 'IMG_100x100' | 'IMG_140x140' | 'IMG_160x160' | 'IMG_170x170' | 'IMG_180x180' | 'IMG_200x200' | 'IMG_312x312' }
```

Figure 2. Example request and response for a specific product, with a flag indicator to show prefetching data.**Request for product:**

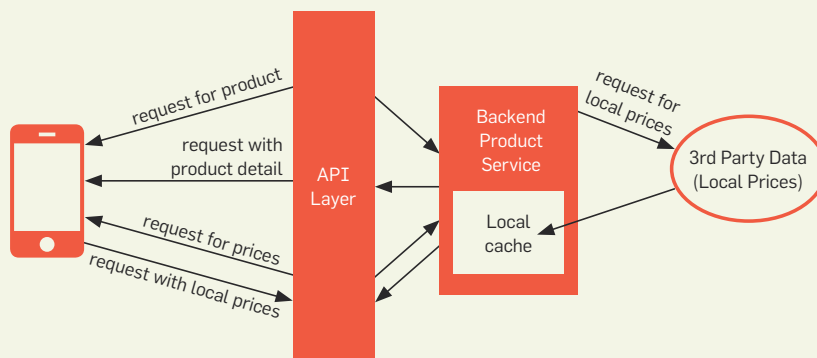
```
http://example.com/api/products.json?productId=18369542&local=true
```

Response:

```
{ "product": {
  "id": "18369542",
  "title": "Upright Freezer",
  "brand": "Frigidaire",
}}
```

Request for prices:

```
http://example.com/api/prices.json?productId=18369542
```

Figure 3. An API call with third-party prefetching (along with Figure 2).

pression, which helps minimize the effect of transmitting additional data.

If using URIs, then make sure to:

- Resize images to the appropriate size before encoding into the URI payload.

- Gzip-compress responses (to take advantage of compression).

- Note that URI-encoded images are part of the CSS of the page. As a result, caching of individual images is more difficult, so do not use this approach if there are good reasons to cache the image locally (that is, it is reused frequently on several pages).

Leverage localStorage and caching.

Since mobile networks can be slow, HTML, CSS, and images can be stored in localStorage to make the mobile experience faster. (There is a great case study on Bing's improvements using localStorage for mobile to reduce the size of an HTML document from about 200KB to about 30KB.¹¹)

Pulling data out of local storage can negatively impact performance,¹³ but it is typically much less than the latency incurred going across the network. In addition to localStorage, some apps are using other features in HTML^{5,6} such as appCache,¹ to improve performance and startup time.

One optimization that can be leveraged on the server involves being aware of what is on the device. By embedding CSS and JavaScript directly within a single Web request, then storing a reference to those files on the client, it is possible to track what has been downloaded and resides in the cache. Then, the next time the client makes a request to the server, it can pass the references to its cached files to the server via a cookie. The server then only has to send new files over the network, which prevents the client from downloading those assets again.

This trick to leverage local caching can save a lot of time. (For more details on how directly to embed and then reference these files, as well as other resources for more reading on the topic, see Mark Pilgrim's *Dive into HTML5*.⁸)

Prefetch and cache data. One great way to improve perceived performance is by prefetching data that will be used throughout the mobile experience so it can be loaded directly on the device without additional requests—for example, paginated results, popular que-

ries, and user data. Thinking about these use cases and factoring them into your API design will allow you to create APIs designed for prefetching and caching data before the user interacts with it, increasing the perception of responsiveness.

If your client is an app, then for data that is not likely to change between updates (such as categories or main navigation) consider shipping the data inside the app so it never requires a trip across the network.

If you want to get sophisticated, ship the data inside the app but also create a versioning and expiration scheme; that way, the app can ping the server in the background and update the data only if the version on the device is out of date.

Ideally, you want to transfer data when needed by the client and preload data when advantageous to do so (such as, when the network or other required resources are not in use). Therefore, if an end user will not view the image or content, then do not send it (this is particularly important for responsive sites since some just “hide” elements). Design your APIs to be flexible and support sending smaller payloads to the client.

A great use case for prefetching images is a gallery of image results, such as a list of products on an e-commerce site. In these situations it is worth downloading the previous and next image(s) to speed up interactions and browsing. Be careful, however, not to go overboard and fetch too far ahead; otherwise, you could end up requesting data that may not be seen by the user.

Use nonblocking I/O. With client optimizations, it is well known to watch out for blocking JavaScript execution,¹⁴ which can have a big impact on the perception of performance. This is even more important for APIs. If there is a longer API call, such as one that could rely on a third party and might time out, it is important to implement this as nonblocking (or even long-waiting) and instead choose a polling or triggering model:

► *Polling API* (pull-based model). The client makes a request and then periodically checks for the results of that request, periodically backing off if required.



You want to make sure that APIs return quickly and do not block while waiting for results, since mobile clients have a limited number of connections.



► *Triggering API* (push-based model). The call makes the request and then listens for a response from the server. The server is provided a call back so it can trigger an event letting the caller know the results are available.

Triggering APIs are typically more difficult to implement, as connections on mobile clients are unreliable. Therefore, polling is a much better option in most cases.

For example, in Decide.com’s mobile app,⁴ each product page shows availability and pricing at stores close to a user’s location. Since a third party delivers those results, the developers did not want the local pricing to take as long as the partner’s API did to deliver results to the client. To work around this, Decide.com created its own wrapper API that allows users to pass a flag for any product query (a set of APIs supported retrieving product data in various ways) that would signal the server to retrieve local prices for that product. Those prices would be stored in the server’s cache. Then in the event the user would want the local pricing for the product, those prices would have a higher probability of being in the cache and would not incur the longer wait times from the third-party partner.

This method is a lot like prefetching on the client but is instead done on the server side with APIs and data. Figure 2 depicts sample requests to show how this works.

As shown in Figure 3, this call looks in the cache first, and if the prices for that product are not present, it calls the third-party API and waits.

In general, you want to make sure that APIs return quickly and do not block while waiting for results, since mobile clients have a limited number of connections. In cases where some components are significantly slower than others on the server side, it can be worth breaking the API into separate calls using typical response time as a factor. That way the client can start rendering pages from the initial fast response calls while waiting for the slower ones. The goal is to minimize the time-to-text rendering on the screen.

You should avoid chatty APIs, and it is important in slow network situations to avoid several API calls. A good rule of thumb is to have all the data

needed to render a page returned in a single API call.

Avoid redirects and minimize DNS lookups. When it comes to requests, redirects can negatively impact performance, especially if they cross domains and require a DNS lookup.

For example, many sites handle their mobile sites using client-side redirects; for example, when a mobile client goes to its main site URL (for example, <http://katemats.com>), it would redirect the client to the mobile site (<http://m.katemats.com>). This is especially common when the sites are built on different technology stacks. Here is an example of how this works:

1. A user Googles “yahoo” and clicks on the first link in the results.
2. Google captures the click using its own tracking URL and then redirects the phone to <http://www.yahoo.com> [redirect].
3. Google’s redirect response goes through the cell tower and then back to the phone.
4. Then there is a DNS lookup for www.yahoo.com.
5. The IP resulting from the DNS lookup is sent through the cell tower and back to the phone.
6. When the phone hits <http://www.yahoo.com>, it is recognized as a mobile client and is redirected to <http://m.yahoo.com> [redirect].
7. The phone then has to do another DNS lookup for that subdomain (<http://m.yahoo.com>).
8. The IP resulting from the DNS lookup is sent through the cell tower and back to the phone.
9. The resulting HTML and assets are finally sent back through the cell tower and then to the phone.
10. Some of the images on pages of the mobile site are served via a CDN (content delivery network), referencing yet another domain, <http://l2.yimg.com>.
11. The phone then has to do another DNS lookup for that subdomain, <http://l2.yimg.com>.
12. The IP resulting from the DNS lookup is sent through the cell tower and back to the phone.
13. The images are rendered, completing the page.

As is obvious from this example, a lot of overhead is involved in these requests. They can be avoided by using redirects on the server side (routing

Design your APIs to allow clients to request just the information they need.

via the server and keeping DNS lookups and redirects to a minimum on the client) or by using responsive techniques.² If DNS lookups are unavoidable, try using DNS prefetching for known domains to save time.

Use HTTP pipelining and SPDY. Another useful technique is HTTP pipelining, which allows for combining multiple requests into one. If I were to implement an optimization translation layer, however, I would opt for SPDY, which essentially optimizes HTTP requests to make them much more efficient. SPDY is getting traction in places such as Amazon’s Kindle browser, Twitter, and Google.

Sending the “Right” Data

Depending on the client, the experience may require different files, CSS, JavaScript, or even the number of results. Creating APIs in a way that supports different permutations and versions of results and files provides the most flexibility for creating amazing client experiences.

Use limit and offset to get results. As with regular APIs, fetching results using `limit` and `offset` allows clients to request ranges of the data that make sense for the client’s use case (thus, fewer results for mobile). The `limit` and `offset` notation is more common (than, say, `start` and `next`), well understood in most databases, and therefore easy to build on:

```
/products?limit=25&offset=75
```

You should choose a default that caters either to the lowest or highest common denominator, depending on which clients are more important to your business: smaller if mobile clients are your biggest users; bigger if users are likely to be on their desktops, such as a B2B website or service.

Support partial response and partial update. Design your APIs to allow clients to request just the information they need. This means that APIs should support a set of fields, instead of returning the full resource representation each time. By avoiding the need for clients to collect and parse unnecessary data, it can simplify the requests and improve performance.

Partial update allows clients to do the same thing with data they are writ-

ing to the API (thereby avoiding the need to specify all elements within the resource taxonomy).

Google supports partial response by adding optional fields in a comma-delimited list as follows:

```
http://www.google.com/calendar/feeds/zachpm@google.com/private/full?fields=entry(title, gd:when)
```

For each call, specifying entry indicates that the caller is requesting only a partial set of fields.

Avoid or minimize cookies. Every time a client sends a request to the domain, it will include all of the cookies that it has from that domain—even duplicated entries or extraneous values. This means keeping cookies small is another way to keep payloads down and performance up. Do not use or require cookies unless necessary. Serve static content that does not require permissions from a cookieless domain, such as images from a static domain or CDN. (The Google Developers site provides some best practices for cookies and performance.⁵)

Establish device profiles for APIs. With the many different screen sizes and resolutions on desktops, tablets, and mobile phones, it is helpful to establish a set of profiles you plan to support. For each profile you can deliver different images, data, and files so they suit each device; you can do this using media queries on the client.¹⁰

If each profile is tailored to a device, then it has the opportunity to offer a better user experience. For each different function and scenario supported by each profile, however, the more difficult it will be to maintain (since devices are constantly changing and evolving). As a result, the smartest approach is to support only as many profiles as absolutely necessary for your particular business. (The mobiForge website offers more information on some trade-offs and options for creating great experiences on different devices.⁹)

For most applications three profiles will be sufficient:

- ▶ Mobile phone—smaller images, touch enabled, and low bandwidth.
- ▶ Tablet—larger images designed for lower bandwidth, touch enabled, more data per request.
- ▶ Desktop—larger, high-resolution

images designed for tablets with high resolution and Wi-Fi or desktop browsers.

Selecting the right profile can be handled by the client, which means on the server side APIs just need to support this configuration. You should design APIs to take these profiles as input, or parameters, and send different information based on the device making the request. Depending on the application, this may mean sending smaller images, fewer results, or inline CSS and JavaScript.

For example, if one of your APIs returns search results to the client, each profile might behave differently as:

```
/products?limit=25&offset=0
```

This would use the default profile (desktop) and serve up the standard page, making a request for each image so subsequent product views could be loaded from cache:

```
/products?profile=mobile&limit=10&offset=0
```

This would return 10 product results and use the low-resolution images encoded as URIs with the same HTTP request:

```
/products?profile=tablet&limit=25&offset=0
```

This would return 20 product results using the larger-size low-resolution images encoded as URIs with the same HTTP request.

You can even create special profiles for devices such as feature phones. Unlike smartphones, feature phones can cache files on only a per-page basis, so it is better to send CSS and JavaScript with each request for these clients. Using profiles is an easy way to support that functionality server side.

You should use profiles instead of partial responses when the response from the server is drastically different per profile—for example, if the response has inline URI images and compact layout in one case but not the other. Of course, profiles could be specified using a “partial response,” although typically it is used to specify a part (or portion) of a standard schema (such as a subset of a larger taxonomy), not a whole different set of data, format, among others.

Conclusion

There are many ways to make the Web faster, including mobile. This article is meant to be a useful reference for API developers who are designing the back-end systems that support mobile clients—and to this end, ultimately enabling and preserving a positive mobile-application user experience. 

Related articles on queue.acm.org

Mobile Application Development: Web vs. Native

Andre Charland and Brian LeRoux

<http://queue.acm.org/detail.cfm?id=1968203>

Streams and Standards: Delivering Mobile Video

Tom Gerstel

<http://queue.acm.org/detail.cfm?id=1066067>

Usability Testing for the Web

Vikram V. Ingleshwar

<http://queue.acm.org/detail.cfm?id=1281891>

References

1. Bidelman, E. A beginner's guide to using the application cache. HTML5 Rocks, 2011; <http://www.html5rocks.com/en/tutorials/appcache/beginner/>.
2. Breheny, R., Jung, E. and Zürrer, M. Responsive design—harnessing the power of media queries. Google Webmaster Central Blog, 2012; <http://googlewebmastercentral.blogspot.com/2012/04/responsive-design-harnessing-power-of.html>.
3. Coyier, C. Which responsive images solution should you use? CS-tricks, 2012; <http://css-tricks.com/which-responsive-images-solution-should-you-use/>.
4. Decide.com. <https://www.decide.com/>.
5. Google Developers. Make the Web faster, 2012; <https://developers.google.com/speed/docs/best-practices/request>.
6. Graham, A. Google APIs + HTML5 = A new era of mobile apps. Google code, 2010; <http://googlecode.blogspot.com/2010/04/google-apis-html5-new-era-of-mobile.html>.
7. Grigsby, J. How Apple.com will serve retina images to new iPads. Cloud Four Blog, 2012; <http://blog.cloudfour.com/how-apple-com-will-serve-retina-images-to-new-ipads/>.
8. Pilgrim, M. The past, present and future of local storage for Web applications. In *Dive into HTML5*, 2009–2011; <http://diveintohtml5.info/storage.html>.
9. Rieger, B. Effective design for multiple screen sizes. mobiForge, 2009; <http://mobiforge.com/designing/story/effective-design-multiple-screen-sizes>.
10. Smus, B. A nonresponsive approach to building cross-device Web apps, 2012; <http://www.html5rocks.com/en/mobile/cross-device/>.
11. Souders, S. Storer case study: Bing, Google, 2011; <http://www.stevesouders.com/blog/2011/03/28/storager-case-study-bing-google/>.
12. W3C Responsive Images Community Group; <http://www.w3.org/community/resimg/>.
13. Zakas, N.C. localStorage read performance. Performance Calendar, 2011; <http://calendar.perfplanet.com/2011/localstorage-read-performance/>.
14. Zakas, N.C. What is a nonblocking script? NCZonline, 2010; <http://www.nczonline.net/blog/2010/08/10/what-is-a-non-blocking-script/>.

Kate Matsudaira is an experienced software engineer and has spent the past seven years immersed in the startup world as an architect or CTO. Prior to that she spent time as a software engineer and technical lead/manager at Amazon and Microsoft. She has a passion for mobile and experience in building large-scale distributed Web systems, cloud computing, and engineering leadership.

© 2013 ACM 0001-0782/13/03

DOI:10.1145/2428556.2428574

The MAGIC 2010 robot competition showed how well multi-robot teams can work with human teams in urban search.

**BY EDWIN OLSON, JOHANNES STROM, ROB GOEDEL,
RYAN MORTON, PRADEEP RANGANATHAN,
AND ANDREW RICHARDSON**

Exploration and Mapping with Autonomous Robot Teams

URBAN RECONNAISSANCE AND search-and-rescue missions are ideal candidates for multi-robot teams due to the potential hazard the missions pose to humans and the inherent parallelism that can be exploited by teams of cooperating robots. However, these domains also involve challenging problems due to having to work in complex, stochastic, and partially observable environments. In particular, non-uniform and cluttered terrain in unknown environments is a challenge for both state-estimation and control, resulting in complicated planning and perception problems. Limited and unreliable communications



further complicate coordination among individual agents and their human operators.

To help address the problems, the Multi-Autonomous Ground robot International Challenge (MAGIC), held November 2010 in Adelaide, Australia, brought together five teams, including nearly 40 robots, to pursue more than \$1 million in prize money in a competition organized and funded by the Australian government's Defence Science and Technology Organisation (DSTO, <http://www.dsto.defence.gov.au/MAGIC2010/>) and the U.S. Army's Research, Development and Engineering Com-



Figure 1. Team Michigan robots. Michigan deployed 14 custom-made robots that cooperatively mapped a 500m × 500m area; each included a color camera and laser range finder capable of producing 3D point clouds.

mand (RDECOM, www.rdecom.army.mil/). The teams were instructed to explore and map a large indoor-outdoor area while recognizing and neutralizing threats (such as simulated bombs and enemy combatants). Although the contest showcased the ability of teams to coordinate autonomous agents in a challenging environment it also reflected the limitations of the state of the art in state estimation and perception (such as map-building and object recognition) (see Figure 1).

MAGIC is the most recent of the robotics Grand Challenges, following in the tradition of well-known com-

petitions sponsored by the Defense Advanced Research Projects Agency (DARPA) tracing back to a 2001 U.S. congressional mandate requiring one-third of all ground combat vehicles to be unmanned by 2015. Over the course of the DARPA challenges, teams developed technologies for fully autonomous cars, including the ability to drive in urban settings, navigating moving obstacles and obeying traffic laws.^{18,19} Moreover, they fostered devel-

» key insights

- Human operators can help a robot team be more efficient and recover from errors.
- A good state estimate, in the form of a map, is the most critical piece of information for a team of robots—and the most difficult to obtain.
- Grand Challenge competitions like MAGIC highlight challenging open problems and provide a venue for evaluating new approaches.

opment of new methods for planning, control, state estimation, and perhaps most important, robot perception and sensor fusion.

Unfortunately, however, these advances were not mirrored in smaller robots (such as those used by soldiers searching for and neutralizing improvised explosive devices, or IEDs) or for robots intended to help first responders in search-and-rescue missions. Instead, tele-operation, or remote-joystick control by a human, remains the dominant mode of interaction (see Figure 2). These real-world systems pose several challenges not in the DARPA challenges:

Limited/unreliable GPS. The Global Positioning System (GPS) is often unreliable or inaccurate in dense urban environments or indoors. GPS can also be jammed or spoofed by an adversary; the winning DARPA vehicles relied extensively on GPS;

Multi-robot cooperation. Robots are individually generally less capable than humans, with their potential

arising from multi-robot deployments that explicitly coordinate with one another; and

Humans in the loop. By allowing humans to interact with robot teams in real time, the system becomes more effective and adaptable to changes in mission objectives and priorities; this ability entails developing visualization methods and user-interface abstractions that allow humans to understand and manipulate the state of the team.

MAGIC focused on increasing the effectiveness of multi-robot systems by increasing the number of robots a single human operator can manage effectively. This is in contrast to more-traditional robot systems that typically require one or more operators per robot.² Participants were required to deploy a team of cooperating robots to explore and map a hostile area, recognize and catalog the location of interesting objects (such as people, doorways, IEDs, and cars), and perform simulated neutralization of IEDs using a laser pointer. Two human operators were allowed to interact with

each robot team, but interaction time was measured and used to calculate a penalty to each team's final score.

The contest attracted 23 teams from around the world, a number reduced through a series of competitive down selects to five finalists invited to the final competition at the Adelaide Showgrounds, a 500m × 500m area including indoor and outdoor spaces. Aerial imagery provided by contest organizers was the only prior knowledge. While previous DARPA challenges provided detailed GPS waypoints describing the location and topology of safe roads, MAGIC robots would have to figure out such information on their own. Whereas other search-and-rescue robotics contests typically focus on smaller environments with significant mobility and manipulation challenges (such as the RoboCup Rescue League, <http://www.robocuprescue.org/>), MAGIC was at a much larger scale, with greater focus on autonomous multi-robot cooperation.¹⁶

To succeed, a team had to combine robot perception, mapping, planning, and human interfaces. Here, we highlight some of the key decisions and algorithmic choices that led to Michigan's first-place finish.¹⁴ Additionally, we highlight how Michigan's mapping and state-estimation system differed from the other competitors, one of the key differences setting Michigan apart.

System Design

The Michigan system was largely centralized: A ground-control station near the center of the competition area collected data from individual robots, fused it to create an estimate of the current state of the system (such as position of robots and location of important objects), then used it to assign new tasks to the robots. Most robots focused on exploring the large competition area, a task well suited to parallelization. However, other robots were able to perform additional tasks (such as neutralizing IEDs). The discovery of such a device would cause a "neutralize" task to be assigned to a nearby robot. Each team's human operators were positioned at the ground-control station where they could view current task assignments, a map of the operating area, and (perhaps most important) guide the system by vetting sensor data

Figure 2. Finalist robots.



(a)



(b)

Each team used a unique robot platform (in ranked order, left to right): Michigan had 14 custom-built robots; Penn had seven custom-built robots; the Reconnaissance and Autonomy for Small Robots team, principally organized by Robotics Research LLC and QinetiQ, based its seven robots on the Talon commercial platform; MAGICian, a coalition of Australian schools, adapted a commercial base for its five robots; and Cappadocia, a coalition based mainly in Turkey, had six custom-built robots.



(c)



(d)

or overriding task assignments.

Michigan's robots received their assignments via radio and were responsible for executing their tasks without additional assistance from the ground-control station; for example, robots used their 3D laser range finder to identify safe terrain and avoid obstacles on their own. They were also responsible for autonomously detecting IEDs and other objects. The information they gathered (including object-detection data and a map of the area immediately around the robot) was heavily compressed and transmitted back to the ground-control station; in practice, these messages were often relayed by other robots to overcome the limited range of Michigan's radios. With the newly collected information, the ground-control station updated its map and user interfaces and computed new (improved) tasks for each robot. This process continued until the mission was complete.

Such a system poses many challenges: How does the ground-control station compute efficient tasks for the robots in a way that maximizes the efficiency of the team? How can humans be kept informed about the state of the system? How can humans contribute to the performance of the system? How do robots reliably recognize safe and unsafe terrain? How do robots detect dangerous objects? How can the information collected by robots be compressed sufficiently so it can be transmitted over a limited, unreliable communications network? And how does the ground-control station combine information from the robots into a single globally consistent view?

Recognizing that many of these tasks rely on an accurate, detailed map of the world, Michigan focused on fusing robot data into a globally consistent view. The accuracy of the map was a primary evaluation criterion in the competition, as well as a critical component in effective multi-agent planning and the human-robot interface; for example, where should robots go next if one does not know where they are now or where they have already been.

A notable difference between Michigan and the other teams was the accuracy of the maps it produced. Map quality pays repeated dividends throughout the Michigan system, with correspond-



While MAGIC posed many technical challenges, mapping and state estimation were arguably most critical.



ing improvement in human-robot interfaces and planning. The variability in map quality from team to team is a testament to the difficulty and unsolved nature of multi-robot mapping. Michigan began with a state-of-the-art system, but it was inadequate in terms of both scaling to large numbers of robots and dealing with the errors that inevitably occur. New methods, both automatic and humans in the loop, were needed to achieve adequate performance; the following section explores a few of them.

Technical Contributions

While MAGIC posed many technical challenges, mapping and state estimation were arguably most critical. Using GPS may seem like an obvious starting point, but even under best-case conditions, it cannot provide a navigation solution for the significant fraction of the time robots are indoors. Outdoors, GPS data (particularly from consumer-grade equipment) is often fairly good, within a few meters, perhaps. GPS can also be wildly inaccurate due to effects like multi-path. In a combat situation, GPS is easily jammed or even spoofed. Consequently, despite having GPS receivers on each robot, Michigan ultimately opted not to use GPS data, relying instead on its robots' sensors to recognize landmarks. This strategy was not universally adopted, however, with most teams using GPS in some way.

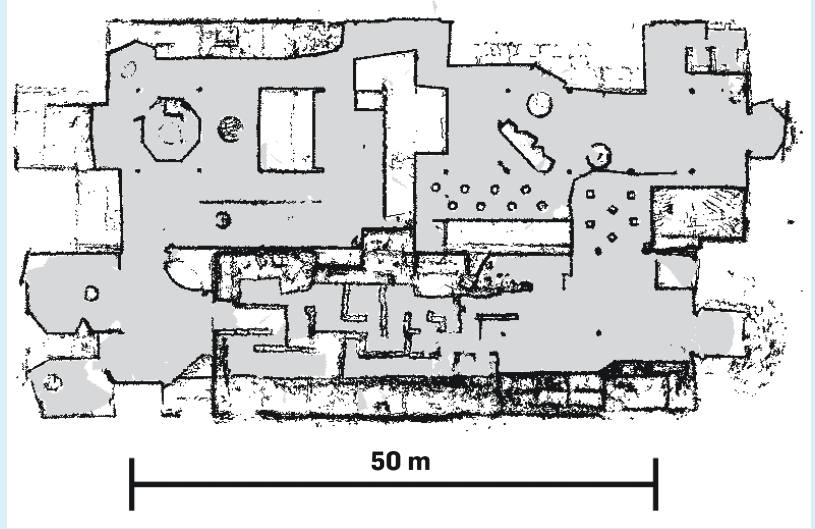
Mapping and state estimation. Conceptually, map building can be viewed as an alignment problem, with robots periodically generating maplets of their immediate surroundings using a laser scanner. The challenge is to determine how to arrange the maplets so they form a large coherent map, much like the process of assembling a panoramic photo from a number of overlapping photos (see Figure 3). Not only can the system assemble a map this way but also the position of each of the robots, since each is at the center of its own maplet.

Michigan's state-estimation system was based on a standard probabilistic formulation of mapping in which the desired alignment can be computed through inference on a factor graph; see Bailey and Durrant-Whyte¹ and Durrant-Whyte and Bailey⁷ for a survey of other approaches. The Michigan

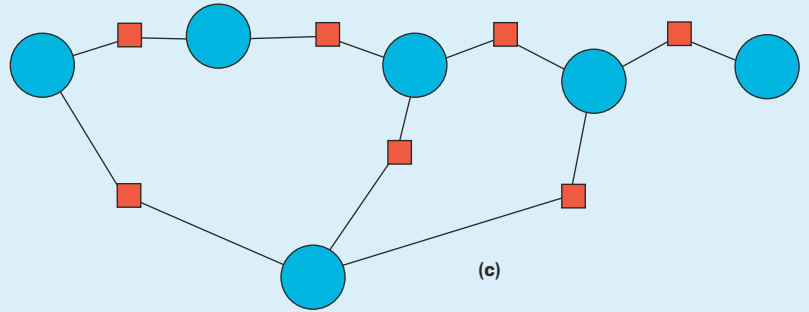
Figure 3. Mapping overview. Individual maplets (a) are matched in a pairwise fashion.



(a)



(b)



(c)

The resulting network of constraints can be represented through a factor graph similar to (c) in which circles represent robot positions and squares probabilistic constraints. The final map (b) is computed by re-projecting all sensor observations according to maximum-likelihood robot positions.

factor graph included nodes for unknown variables—the location of each maplet—and edges connecting nodes when something is known about the relative geometric position of the two nodes. Loosely speaking, an edge encodes a geometrical relationship between two maplets; that is “maplet A is six meters east and rotated 30 degrees from maplet B”; none of these relationships is known with certainty, so edges are annotated with a covariance matrix. A map commonly contains many of these edges, with many of them subtly disagreeing with one another.

More formally, let the position of all maplets be represented by the state vector x , which can be quite large, as it contains two translation and one rotation component for each maplet, and there can be thousands of maplets. Edges convey a conditional probability distribution $p(z_i|x)$, where z_i is a sensor measurement. This quantity is the measurement model; given a

particular configuration of the world, it predicts the distribution of the sensor; for example, a range sensor might return the distance between two variable nodes plus some Gaussian noise the variance of which can be characterized empirically.

Our goal was to compute $p(x|z)$, or the posterior distribution of the maplet positions, given all sensor observations. Using Bayes’s rule, and assuming we have no a priori knowledge of what the map should look like (or $p(x)$ is uninformative), we obtain:

$$p(x|z) \propto \prod p(z_i|x) \quad (1)$$

The goal was to find the maplet positions x that has maximum probability $p(x|z)$. Assuming all the edges are simple Gaussian distributions of the form $e^{(z_i - \mu)^T \Sigma^{-1} (z_i - \mu)}$, this computation becomes a nonlinear least-squares problem. Specifically, we can take the logarithm of both sides, which con-

verts the right-hand side into a sum of quadratic losses. We maximize the log probability by differentiating with respect to x , resulting in a first-order linear system. The key idea is that maximum likelihood inference on a Gaussian factor graph is equivalent to solving a large linear system; see Thrun et al.¹⁷ for a more detailed explanation. The solution to this linear system yields the position of each maplet.

The resulting linear system is extremely sparse, as each edge typically depends on only two maplet positions. In the Michigan system, each maplet was generally connected to from two to five other maplets. Sparse linear algebra methods can exploit this sparsity, greatly reducing the time needed to solve the linear system for x . The Michigan method was based on sparse Cholesky factorization;⁶ we could compute solutions for a graph with 4,200 nodes and 6,300 edges in about 250ms on a standard laptop CPU. New data is

always arriving, so this level of performance allowed the map to be updated several times per second.

An important advantage of using the factor graph formulation is that it is possible to retroactively edit the graph to correct errors; for example, if a sensing subsystem erroneously adds an edge to the graph (incorrectly asserting that, say, two robot poses are a meter apart), we could “undo” the error by deleting the edge and computing a new maximum likelihood estimate. Such editing is not possible with methods based on, say, Kalman filters. In this case, we relied on human operators to correct these relatively rare errors, discussed later.

Scan matching and loop validation.

The Michigan mapping approach depended on identifying high-quality edges; more edges generally result in a better map since the linear system becomes over-constrained, reducing the effect of noise from individual edges.

The system used several different methods to generate edges, including dead reckoning (based on wheel-encoder odometry and a low-cost inertial measurement unit, or the set of sensors that measures acceleration and rotation of the robot) and visual detection of other robots using their 2D “bar codes,” as in Figure 1.¹⁰ But the most important source of edges in the Michigan system (by far) was its scan-matching system, attempting to align two maplets by correlating them against each other, looking for the translation and rotation that maximize their overlap. One such matching operation (see Figure 4) includes the probability associated with each translation and rotation computed in a brute-force fashion.

This alignment process is computationally expensive, and in the worst case, each maplet had to be matched with every other maplet. In practice, the robot’s dead-reckoning data can help rule out many false matches. But with 14 robots operating simultaneously, and with each one producing a new maplet every 1.4 seconds, hundreds or thousands of alignment attempts per second are needed.

The Michigan approach to mapping was based on an accelerated version of a brute-force scan-matching system.¹¹ The key idea is a multi-resolution matching system, generating low-res-

olution versions of the maplets and a first attempt to align them. Because they are smaller, the alignment is much faster. Good candidate alignments are then attempted at higher resolution.

While simple in concept, a major challenge was ensuring the low-resolution alignments did not underestimate the quality of an alignment that could occur with higher-resolution maplets. The Michigan solution relied on constructing the low-resolution maplets in a special way; rather than apply a typical low-pass-filter/decimate process (which would tend to obliterate structural details), we used a max-decimate kernel to ensure matches between low-resolution maplets never underestimate the overlap that could result from aligning full-resolution maplets. When aligning low-resolution maplets, we never underestimated the overlap that could result from aligning the full-resolution maplets.

Michigan’s earlier scan-matching work¹¹ considered two different resolutions, allowing approximately 50 matches per second. This would be adequate for a small number of robots, but for a larger team, it becomes a bottleneck. For MAGIC, we modified the approach to consider matches over a full pyramid of reduced-resolution images that resulted in matching rates of approximately 500 matches per second. The quality of the resulting map ultimately depended on the number of matches found, and a higher processing rate increases the likelihood

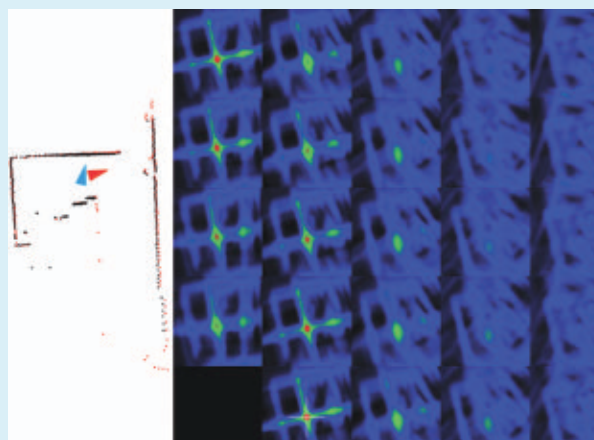
of finding these matches. Our fast-matching system was pivotal in keeping up with our large robot team. Other teams used similar maplet-matching strategies though were not as fast; for example, the Australian MAGICian team reported its GPU-accelerated system was capable of seven to 10 matches per second.

The improvement in our matching speed allowed us to consider a large number of possible matches in real time to support our global map. However, our state-of-the-art method had a non-zero false-positive rate, aligning maplets based on similar-looking structures, even if the maplets are not actually near each other.

There is a fundamental trade-off between number of true positives and increased risk of false positives. Increasing the threshold for what constitutes a “good-enough” match also increases the likelihood that similar looking, but physically distinct, locations are matched incorrectly. Such false-positive matches can cause the inference method to distort the map to explain the error.

To reduce the false-positive rate to a usable level, we performed a loop-validation step on candidate matches before the system could add them to the factor graph. The basic idea of loop validation is to require that multiple matches “agree” with each other.^{4,12,13} Consider a topological “loop” of matches: A match between node A and B, another match between B and

Figure 4. Brute-force search for best maplet alignment. The search space is 3D (two translation, one rotation) represented as a series of 2D cross-sections.



Bright areas indicate good alignment; finding the best match quickly is critical in large-scale mapping systems, with resulting matches becoming edges in the factor graph.

C, and a third match between C and A. If the matches are correct, then the composition of their rigid-body transformations should approximately be the identity matrix. When this occurs, the system adds the matches to the factor graph.

Human-robot interfaces. In simple environments (such as an indoor warehouse), the combination of loop-validation and automatic scan-matching presented here were sufficient for supporting completely autonomous operation of Michigan's robot team (see Figure 5). However, in a less-structured environment (such as many of the outdoor portions of MAGIC 2010), mapping errors would still occur; for example, the MAGIC venue included numerous cable conduits that caused robots to unknowingly get stuck, causing severe dead-reckoning estimation error. At the time of MAGIC 2010, Michigan's system was not able to handle such problems autonomously.

However, map errors are relatively obvious to human operators. Michigan thus developed a user interface that allowed human operators to look for

errors and intervene when necessary. With new (validated) loop closures being added to the graph at a rate of two to three per second, a human operator could easily be overwhelmed by asking for explicit verification of each match.

Instead, human operators would monitor the entire map. When an error occurred (typically visible as a distortion in the map), the operator could "roll back" automatically added matches until the problem was no longer present. The operator could then ask the mapping system to perform an alignment between two selected maplets near where the problem had been detected. This human-assisted match served as additional a priori information for future autonomous matching operations, making it less likely the system would repeat the same mistake.

Michigan found this approach, which required only a few limited interactions to remove false positives, a highly effective use of humans to support the continued autonomy of its robots' planning system. Michigan was the only team to build a user interface that allowed direct supervision of the real-time state estimate; other teams

handled failures in automatic state estimation by requiring humans to track the global state manually, then intervene at the task-allocation level. Early versions of the system lacked a global-mapping system, with human operators providing separate map displays for each robot. Michigan's experience with this approach indicated that operators could not effectively handle more than five or six robots this way. Maintaining a global map is critical to scaling to larger robot teams, and the Michigan user interface was a key part of maintaining map consistency.

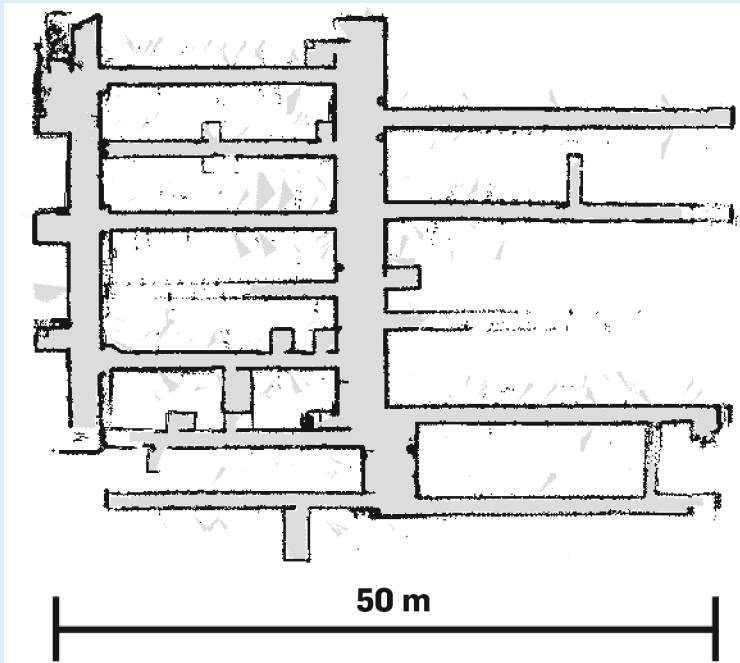
Evaluation

The main evaluation metrics for an autonomous reconnaissance system are quality of the final map produced and amount of human assistance required to produce it and were also the primary metrics the MAGIC organizers used to determine the winner and subsequent ranking of the finalists, as in Figure 2. While the specific performance data used in the competition was not made public, we present selected results we obtained by processing our logs; we also compare with other teams' published results where possible.

Lacking detailed ground truth for the MAGIC venue, the best evaluation of map quality is necessarily subjective; Figure 6 compares post-processed maps for the Michigan team against the mapping software of MAGICian (fourth place) applied to the data collected by the Penn team (second place); additionally, the map produced by the Michigan system (inset), includes distortions resulting from erroneous matches that, in the interest of time, the human operators chose not to correct. This result shows that high-quality maps can be produced in this domain; Michigan's competition-day results showed our state estimation was sufficiently good to be useful for supporting online planning. The system allowed us to completely explore the first two phases of the competition while simultaneously performing mission objectives relating to dynamic and static dangers (such as IEDs and simulated mobile enemy combatants).

We would also would like to measure the frequency of human inter-

Figure 5. Indoor-storage-warehouse map. In uncluttered environments posing few mobility challenges, Michigan's team of 14 robots could explore and map with little human intervention.



action required to support our state estimation system during MAGIC. However, Michigan did not collect the data necessary to evaluate this metric during our run; we thus replicated the run by playing back the raw data from the competition log and having our operator reenact his performance during the competition. These conditions are obviously less stressful than competition but are still representative of human performance. The result (see Figure 7) was the addition of 175 loop closures, on average two interactions per minute, which generally occurred in bursts. However, at one point, the operator did not interact with the system for 5.17 minutes.

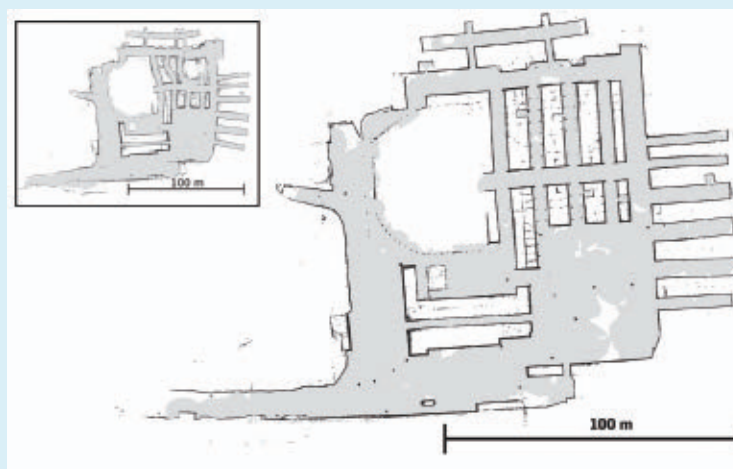
Our evaluation shows we were able to support cooperative global state estimation for a team of autonomously navigating robots using only a single part-time operator. Yet there remain significant open problems, including how to reduce human assistance even further by improving the ability of the system to handle errors autonomously. Additional evaluation of the system, as well as technical descriptions of the other finalists, can be found elsewhere, including in Boeing et al.,³ Butzke et al.,⁵ Erdener,⁸ Lacaze et al.,⁹ and Olson et al.¹⁴

Discussion

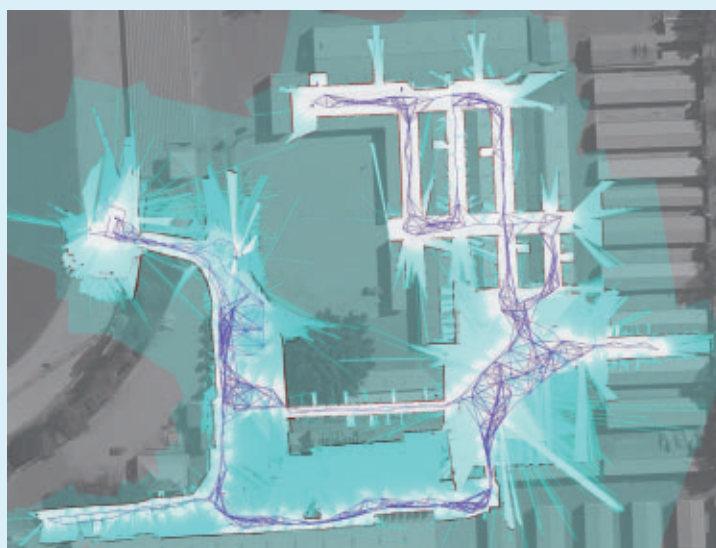
MAGIC's focus was on increasing the robot-to-human ratio and efficiently coordinating the actions of multiple robots. Key to reducing the cognitive load on operators is how to increase the autonomy of robots; for a given amount of cognitive loading, more robots can be handled if they simply require less interaction. We identified global state estimation as a key technology for enabling autonomy and believe the mapping system we deployed for MAGIC outperforms the systems of our competitors. While this was one of the key factors differentiating Michigan from the other finalists, it was not the only important point of comparison. In fact, many of the other choices we made when developing the system also had a positive effect on our performance.

In particular, we made a strategic decision early on that we would emphasize a large team of robots. This is reflected in the fact that we brought

Figure 6. Minimally post-processed maps from Michigan's robots (a) and MAGICian's mapping algorithm using Penn's data (b) from Reid and Brauni.¹⁵ The map produced online by the Michigan robots is inset top-left.



(a)



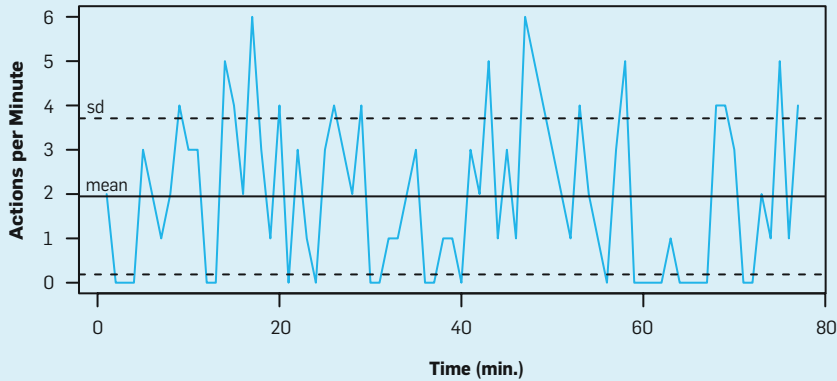
(b)

twice as many robots to the competition as the next largest team. This strategy ultimately affected the design of all our core systems, including mapping, object identification, and communication. Given that we had a finite budget, it also forced us to deploy economical robot platforms with only the bare necessities in sensing to complete the challenge. The result was our robots were also the cheapest of any of the finalists (by a significant margin), costing only \$11,500 each.

One approach to detecting dangerous objects is to, say, transmit video feeds back to human operators and

rely on them to recognize the hazard. Given a design goal of maximizing the number of robots, such a strategy is unworkable; there is neither the bandwidth to transmit that many images nor could humans be expected to vigilantly monitor 14 video streams. The system simply had to be able to detect dangerous objects autonomously, whereas other teams with fewer robots could be successful with less automation. At the same time, handling more tasks autonomously also meant our human operators had more time to assist with mapping tasks.

Figure 7. Map interaction experiment. Michigan's mapping operator reenacted the supporting role for the phase 2 dataset to measure the frequency of interaction required to maintain a near-perfect state estimate; see Figure 6 for resulting map. The human workload was modest, averaging only two interactions per minute.



An interesting question relates to the optimal size of the robot team; larger teams have greater parallelism-exploiting potential but also place greater demands on human operators. These demands are strongly dependent on the reliability and autonomous capabilities of the robots; less-capable robots would place greater demands on the operators. An area of future work involves trade-offs between the size of a robot team and the cognitive load on human operators and how the autonomous capabilities of the robots affect this trade-off.

Conclusion

MAGIC resulted in significant progress toward urban search using teams of robots aided by human operators. Michigan's approach, emphasizing accurate mapping, helped maximize the autonomous capabilities of its robots and maintain the operator's situational awareness, allowing two humans to effectively control a larger team of robots. However, MAGIC also highlighted the shortcomings of state-of-the-art methods. It remains difficult to maintain a consistent map for large numbers of robots; Michigan's competition-day maps still show distortions due to errors in the system's matching capability. The system coped with these errors at the cost of greater operator workload we continue to target in our ongoing work.

Competitions like MAGIC highlight open technological challenges in areas often viewed as "solved." MAGIC thus

brought the prospect of cooperative teams of robots and humans closer than ever, but also highlighted the challenging research problems that remain.

Acknowledgments

Team Michigan was a collaboration between the University of Michigan's APRIL Robotics Laboratory (<http://april.eecs.umich.edu/>) and Soar Technology (<http://www.soartech.com/>). In addition to the authors here, core team members included Mihai Bulic, Jacob Crossman, and Bob Mariner; we were also supported by more than two dozen undergraduate researchers. We thank the MAGIC contest organizers who ran the competition and prepared the contest venue. And special thanks go to our liaison, Captain Chris Latham of the 9th Combat Service Support Battalion of the Australian Army. Our participation would not have been possible without the help of our sponsors at Intel and Texas Instruments. 

References

1. Bailey, T. and Durrant-Whyte, H. Simultaneous localization and mapping (SLAM): Part II state of the art. *Robotics and Autonomous Systems* 13, 3 (Sept. 2006), 108–117.
2. Barnes, M. and Jentsch, F. *Human-Robot Interactions in Future Military Operations*. Ashgate Publishing Company, Brookfield, VT, 2010.
3. Boeing, A., Boulton, M., Brun, T., Frisch, B., Lopes, S., Morgan, A., Ophelders, F., Pangen, S., Reid, R., and Vinsen, K. WAMbot: Team MAGICian's entry in the Multi-Autonomous Ground-Robotic International Challenge 2010. *Journal of Field Robotics* 5 (Sept./Oct. 2012), 707–728.
4. Bosse, M.C. *ATLAS: A Framework for Large-Scale Automated Mapping and Localization*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, Feb. 2004.
5. Butzke, J., Danilidis, K., Kushleyev, A., Lee, D.D., Likhachev, M., Phillips, C., and Phillips, M. The University of Pennsylvania MAGIC 2010 Multi-Robot

- Team. *Journal of Field Robotics* 5 (Sept./Oct. 2012), 745–761.
6. Dellaert, F. and Kaess, M. Square root SAM: Simultaneous localization and mapping via square root information smoothing. *International Journal of Robotics Research* 25, 12 (Dec. 2006), 1181–1203.
7. Durrant-Whyte, H. and Bailey, T. Simultaneous localization and mapping (SLAM): Part I, the essential algorithms. *Robotics and Autonomous Systems* 13, 2 (June 2006), 99–110.
8. Erdener, A., Ari, E.O., Ataseven, Y., Deniz, B., Ince, K.G., Kazancioglu, U., Kopanoglu, T.A., Koray, T., Kosaner, K.M., Ozgur, A., Ozkok, C.C., Soncul, T., Sirin, H.O., Yakin, I., Biddlestone, S., Fu, L., Kurt, A., Ozguner, U., Redmill, K., Aytekin, O., and Ulusoy, I. Team Cappadocia design for MAGIC 2010. In *Proceedings of the Land Warfare Conference* (Brisbane, Australia, Nov. 15–19, 2010).
9. Lacaze, A., Murphy, K., Giorno, M.D., and Corley, K. The Reconnaissance and Autonomy for Small Robots (RASR) MAGIC 2010 Challenge. In *Proceedings of the Land Warfare Conference* (Brisbane, Australia, Nov. 15–19, 2010).
10. Olson, E. AprilTag: A robust and flexible visual fiducial system. In *Proceedings of the IEEE International Conference on Robotics and Automation* (May). IEEE, 2011, 3400–3407.
11. Olson, E. Real-time correlative scan matching. In *Proceedings of the IEEE International Conference on Robotics and Automation* (Kobe, Japan, June). IEEE, 2009, 4387–4393.
12. Olson, E. Recognizing places using spectrally clustered local matches. *Robotics and Autonomous Systems* 57, 12 (Dec. 2009), 1157–1172.
13. Olson, E. *Robust and Efficient Robotic Mapping*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, June 2008; <http://april.eecs.umich.edu/papers/details.php?name=olson2008phd>
14. Olson, E., Strom, J., Morton, R., Richardson, A., Ranganathan, P., Goeddel, R., and Bulic, M. Progress towards multi-robot reconnaissance and the MAGIC 2010 competition. *Journal of Field Robotics* 29, 5 (Sept. 2012), 762–792.
15. Reid, R. and Braun, T. Large-scale multi-robot mapping in MAGIC 2010. In *Proceedings of the IEEE Conference on Robotics, Automation, and Mechatronics* (Sept. 17–19). IEEE, 2011, 239–244.
16. Saenbunsiri, K., Chaimuegchuen, P., Changlor, N., Skolapak, P., Danwang, N., Ppoosuwann, V., Tienkum, R., Anan Raktrajulthum, P., Nitisuchakul, T., Bumrungritt, K., Tunsiri, S., Khairid, P., Santi, N., and Yan Primee, S. *RobotCupRescue 2011: Robot League Team iRAP JUDY (Thailand)*. Technical Report, 2011.
17. Thrun, S., Burgard, W., and Fox, D. *Probabilistic Robotics*. MIT Press, Cambridge, MA, 2005.
18. Thrun, S. et al. Stanley: The robot that won the DARPA Grand Challenge. In *The 2005 DARPA Grand Challenge, Vol. 36 of Springer Tracts in Advanced Robotics*. Springer Berlin/Heidelberg, 2007, 1–43.
19. Urmson, C. et al. Autonomous driving in urban environments: Boss and the urban challenge. *Journal of Field Robotics* 25, 8 (Aug. 2008), 425–466.

Edwin Olson (ebolson@umich.edu) is an assistant professor in the Department of Computer Science and Engineering at the University of Michigan, Ann Arbor.

Johannes Strom (jstrom@umich.edu) is a Ph.D. student focusing on communications and mapping in the Department of Computer Science and Engineering of the University of Michigan, Ann Arbor.

Rob Goeddel (rgooddel@umich.edu) is a Ph.D. candidate focusing on exploration and planning in the Department of Computer Science and Engineering of the University of Michigan, Ann Arbor.

Ryan Morton (rmorton@umich.edu) is a Ph.D. candidate focusing on terrain classification and human interaction in the Department of Computer Science and Engineering of the University of Michigan, Ann Arbor.

Pradeep Ranganathan (pradeep@umich.edu) is a Ph.D. student focusing on learning and inference in the Department of Computer Science and Engineering of the University of Michigan, Ann Arbor.

Andrew Richardson (arichardson@umich.edu) is a Ph.D. candidate focusing on computer vision in the Department of Computer Science and Engineering of the University of Michigan, Ann Arbor.

They deliver the right social service to the right user anytime, anyplace, without divulging personal data.

BY NAFAÂ JABEUR, SHERALI ZEADALLY, AND BIJU SAYED

Mobile Social Networking Applications

RECENT ADVANCES IN mobile computing, hardware, and software empower end users worldwide through a range of mobile devices (such as smartphones and tablets) with improved and novel capabilities (such as localization through the Global Positioning System, context awareness through sensors, Internet access

through cellular networks, and short-range communications through Wi-Fi networks). The result is intense competition among providers of online social services for mobile users regardless of location and profile, along with numerous mobile social networking (MSN) applications in which billions of people use their mobile devices to tap a spectrum of instant, relevant, high-quality services (such as interaction with peers with similar interests, sharing information, and creating virtual communities).

» key insights

- **Along with PC functions, TV, games, and business services are available through mobile devices, wherever and whenever a user might want them.**
- **Mobile devices are readily discoverable by nearby people and social services.**
- **Even more services are expected soon, along with numerous challenges and questions about privacy and data security.**

Like online social networking sites (OSNS), MSN applications are social structures consisting of individuals or organizations connected through specific types of interdependency (such as friendship, kinship, common interest, financial exchange, and beliefs). They are based on a variety of architectures depending on whether they are extensions of existing OSNS, designed purely for mobile devices, focused on mobile users, or data- or service-oriented. Services available to mobile users follow several trends, including social gaming, business, and media. To better understand the state of MSN applications, we reviewed their architectures, trends, and impact over the past few years, motivated by the lack of previous studies surveying MSN applications in both the business sector and the research community. We were also motivated by our strong interest in understanding how MSN

applications provide services to the right users at the right time through communication and context-aware technologies. One of our major contributions is proposed improvements that can be incorporated into existing MSN services, enabling seamless migration from PC-based environments to mobile environments. In addition, we also provide broad insight into state-of-the-art MSN applications, identifying strengths and weaknesses, classifications, and a new proposed classification under which a non-exhaustive set of MSN applications can be identified.

Mobile Social Networks

Several surveys previously revealed the dramatic increase in popularity of MSN applications. Indeed, a 2011 survey found 53% of mobile users in North America used these applications.⁸ Another survey found nearly 40% (almost 300 million people worldwide) of those accessing monthly social networks from their mobile devices are Facebook users.²³ The analytics firm ComScore³ reported the number of people accessing social networks from their phones in France, Germany, Italy, Spain, and U.K. grew by 44% in 2011, reaching 55 million users in September 2011 in these five countries. ComScore also reported that Twitter and LinkedIn more than doubled their numbers of users in the same year. Meanwhile, Warren²⁵ concluded that the number of mobile subscribers accessing Facebook and Twitter increased by 112% and 347% from January 2009 to January 2010, respectively, while the number decreased by 7% for MySpace in the same period (see Table 1).

The increasing popularity of MSN applications is notably due to new, interesting services and ways of engaging in social interaction and collaboration through mobile devices. Indeed, in addition to locating and alerting users about friends and communities, users can also use location-based services (such as to recommend nearby commercial offers) and data-sharing services (such as photos). Some services have been extended from PC-based social networking sites to be available almost everywhere, anytime, for mobile devices following their location- and proximity-aware facilities (see Figure 1).



In addition to user profiles, MSN applications should address users' emotional states, as well as the focus of their attention.



Several MSN initiatives have been proposed; for example, Smith²¹ presented the Reno mobile-phone application in which users query one another and exchange location information in response to other queries or even when unprompted. In Reno, mobile devices are classified into types and matched to specific types of queries. The Reality Mining project⁶ demonstrates the ability of Bluetooth-enabled mobile phones to measure information access in different contexts, recognize social patterns in routine user activity, infer relationships, and identify socially significant locations. The CenceMe system¹⁵ collects users' status or context information through mobile sensors, exporting it automatically to social networks. Serendipity⁵ uses Bluetooth to find nearby devices and a central server to match profiles for either a professional introduction or for more personal reasons (such as dating). Nicolai et al.¹⁶ proposed an application that relies on neighboring-device discovery to sense and visualize the surrounding social network on mobile devices. MobiClique¹⁸ is a mobile ad hoc network in which Bluetooth-enabled mobile devices communicate directly with other devices as they meet opportunistically. The CityWare system described in Kostakos and Neill¹⁰ is built on a similar idea but passes proximity information to an online social-networking application that aggregates and sends statistics about users' surroundings.

From PC- to Mobile-based Environment

The technological progress in mobile devices, communication facilities, and context-aware capabilities is the major driver behind the shift among social-networking sites from Web-based to hybrid to pure mobile applications. Mobile applications attract the attention of mobile users who want social services anytime, anywhere, eliminating the need for desktop PCs. In contrast to PC-based social-networking sites, important features for users in mobile social services include immediacy, relevance, brevity, and retrieval.⁷ Immediacy means users get answers to their questions or report on important events on the fly. Relevance means MSN applications use location-aware devices to send queries and messages

Figure 1. MSN services and their providers; provider key: N: network; A: application; S: service; Y: system.

Main Idea	Providers and Links
Services based on location-aware devices	
The MSN provider gets the locations of its users from their GPS-enabled mobile devices, using the information to locate nearby friends, provide recommendations, and allow users to discover their surroundings.	Aka-Aki (A) (http://www.aka-aki.com/) Brightkite (S) Foursquare (S) (https://foursquare.com/) Gowalla (N) (http://gowalla.com/) Loopt (A) (https://www.loopt.com/) Playtxt (N) (http://www.playtxt.net) Plazes (Y) (http://www.plazes.com)
Services based on proximity-aware devices	
The MSN provider allows its users to use their Bluetooth-enabled devices (such as Aki-Aki, Bluedating, Lovegety, MobiClique, Nokia Sensor, Proximating, Speck, and Tothing) or Wi-Fi connectivity (such as Aka-Aki, FaceTime, and Jambo) to find and communicate with nearby friends and others with similar profiles and interests.	Bluedating (S) (http://www.bluedating.com/) FaceTime (S) (http://www.apple.com/iphone/built-in-apps/facetime.html) Jambo (N) (www.jambo.net) LoveGety (Y) Nokia Sensor (A) Proximating (S) (http://www.proximating.com) Speck (Y) (http://speck.randomfoo.net/) Tothing (S)
Services provided by centralized servers	
The MSN provider has a remote server with which mobile users interact to get services, including recommendations (such as Whrrl and Yelp), finding friends (such as MobiLuck), seeking and exchanging information and goods (such as PeopleNet), exchanging messages, viewing profiles, reading and sending bulletins, and viewing photos (such as Bebo, Facebook, Myspace, MyYearBook, and Twitter), downloading and playing games (such as Friendster and Zynga), streaming video (such as YouTube), making professional contacts (such as LinkedIn), and learning (such as iTeach).	Bebo (S) (http://www.bebo.com/m/) Facebook (S) (http://www.facebook.com/mobile/) Friendster (S) (http://m.friendster.com/) iTeach (A) (http://grou.ps/iteachmobile) LinkedIn (S) (http://touch.www.linkedin.com/mobile.html) MobiLuck (S) (http://www.mobiluck.com/en/) Myspace (S) (http://www.myspace.com/) MyYearBook (S) (http://www.myearbook.com/mobile.php) PeopleNet (N) (www.peoplenetonline.com/) Twitter (S) (http://twitter.com/#!/twittermobile) Whrrl (S) Yelp (S) (http://www.yelp.com/) YouTube (S) (http://www.youtube.com/mobile) Zynga (A) (http://company.zynga.com/games/mobile-games)

to people within a defined geographic area, enabling groups of users to share an experience virtually. Brevity means short messages delivered through mobile devices are easier for others to understand and respond to. And retrieval means conversations are archived and retrieved later by participants or others, creating a kind of real-time archive of social interactions.

To reap these benefits, MSNs are adapting the way they provide their services, especially by minimizing explicit user intervention while aiming to deliver the right content to the right user at the right time. To this end, in addition to providing similar PC-based social-networking services, MSNs must be able to capture optimistically relevant contextual features in users' surroundings.¹³ Such features include location-related, user-related, device-related, interaction-related, and spatio-tempo-related attributes.

Location-related attributes. Several location-related attributes may be re-

levant to MSN applications, depending on the service to be provided to the user, including type of location (such as public space, restaurant, and classroom) and neighboring objects of interest at that location (such as friends and others with similar profiles and landmarks recommended by friends). To capture the characteristics related to user location, MSN applications can benefit from such technologies as GPS, sensors, radio frequency identification (RFID), and near-field communication (NFC) chips in smartphones to establish radio communication by touching

them together or even by just bringing them into close proximity. Several commercial MSN providers, including Aka-Aki, BrightKite, Dodgeball, Mobiluck, and Plazes, use location-aware devices. Several research efforts have also proposed MSN applications based on GPS, including Marmasse,¹⁴ and/or proximity sensing, including Eagle and Pentland,⁶ Kostakos and Neill,¹⁰ Miluzzo et al.,¹⁵ and Nicolai et al.¹⁶ CenceMe¹⁶ is an example of a system that takes inputs from a broad set of sensors, automatically learns from each user's history of digital behavior,

Table 1. Number of mobile subscribers accessing specific social networking sites via mobile browsers in 2010; source: C. Warren²⁵

	Jan. 2009 (000s)	Jan. 2010 (000s)	% change
Facebook	11,874	25,137	+112%
MySpace	12,338	11,439	-7%
Twitter	1,051	4,700	+347%

and outputs status information much richer than current location and communication preference. CenceMe's ability to learn is important for MSN applications in which implicit information is inferred from GPS and sensor data to recommend nearby spatial objects (such as landmarks) likely to be of interest to the user. Likewise, an MSN application should be able to identify nearby people who might be of interest to the user (such as friends or friends of friends within the same geographical area).

User-related attributes. As the user is the focus of MSNs, several commercial applications deliver services based on personal profiles; for example, Loopt, Mobiluck, Playtxt, and Proximating all basically compare user profiles and preferences, sending an alert when a positive match is confirmed. In addition to user profiles, MSN applications should address users' emotional states, as well as the focus of their attention. Also helpful is for them to identify and automatically update users' status (such as participation in sports and work or relaxing at home).⁹ To the best of our knowledge, commercial MSN applications do not yet support such computationally complex issues. Among researchers, inference of user activity is addressed through various approaches; for example, in SenSay,²⁰ a smartphone prototype takes advantage of user context to improve usability, so if the user is busy and wishes to not be disturbed, the smartphone can answer/reply automatically through a short message service. Marmasse¹⁴ developed a system that uses GPS data, accelerometer data to distinguish between walking and driving, and a microphone to distinguish between talking and silence.

Device-related attributes. Mobile devices are characterized by processing, memory, sensing, and battery capability, as well as display screen and compatibility with existing technology. These parameters are being improved on such devices as smartphones and tablets; for example, smartphones, which may contain a large number of sensors and integrated devices, are being upgraded into powerful computing platforms. However, such progress is not al-

ways accompanied by corresponding progress in MSN services. Indeed, existing MSN applications are not always available on all platforms and devices; for example, Jambo and Toothing are available for cellphones and PDAs, Whrrl can be downloaded onto the BlackBerry Pearl, Curve, and Nokia N95 smartphones, and Friendster is optimized for Android, iOS, and Windows Phone 7 smartphones with screens larger than 3.5 inches. For convenience, MSN providers support services that comply with current standards, as compliance yields better interoperability, particularly among mobile devices supported by different technologies. Compliance also yields enhanced games and social-media services because many of them have specific processing and display requirements.

Spatio-temporal-related attributes. Spatio-temporal events are important features of context awareness. Events like day, night, and rain have different effects on users, as well on the services provided to them through mobile devices. Indeed, users could be disturbed by nearby events (such as heavy rain or loud sounds like thunder) so their moods might influence the service they are looking for and the way they interact with others through MSN applications. Despite the importance of these effects, few reported research efforts (such as Liaquat et al.¹¹) address spatio-temporal attributes and their effect on MSN applications. Such efforts have focused on the behavior of mobile users while using their devices over various time periods; they have also studied temporal social communications where different centrality measures (such as proximity) can help determine optimal ways to disseminate information within social networks. Further work is needed to capture and analyze spatio-temporal events and their effects on MSN services. Sensors may help, as they are promising tools for data acquisition and for capturing patterns of user behavior during spatio-temporal events.

Social-interaction-related attributes. MSNs offer a novel user-interaction paradigm combining the benefits of PC-based social networks and mobile-computing devices.²² This interaction might be achieved by sending and receiving

text messages (such as SMS), multimedia messaging service, or MMS, and/or email. As these facilities are not necessarily available on mobile devices, interaction among devices is not always straightforward. In commercial MSNs, device interaction is achieved basically through Bluetooth (such as in Aka-Aki), Wi-Fi connectivity (such as in Jambo), and mobile Internet connections (such as in Facebook and Twitter). Among researchers, BlueFriend²² takes advantage of mobile devices and Bluetooth technology to scan the environment for members of the BlueFriend community. If found, virtual personal cards (with user profiles and preferences) are exchanged to assess the degree of matching among nearby users.

Serendipity⁵ combines the existing communication infrastructure with online-introduction-system functionality to facilitate interaction between physically close people through a centralized server. It repeatedly scans for Bluetooth devices, transmitting the discovered devices to a server that calculates a similarity score between any two proximate users. When this score reaches a predefined threshold, the server alerts both users, sending them information that might include pictures, news of mutual interest, and talking points.

Architectural Considerations

MSN follows three different types of architecture: centralized (such as Facebook and Twitter), peer-to-peer, or P2P (such as BlackBerry Messenger, Lovegety, and Proximating), and hybrid (such as Jambo). Centralized architectures allow users to access multiple services by interacting with remote MSN servers through their mobile devices, freeing the devices from overly demanding processing load and extending battery lifetimes. P2P architectures allow users to interact directly through specific software tools and hardware facilities (such as Bluetooth and Wi-Fi) on their devices. In addition to sharing similar contexts, users may also meet face to face when in neighboring locations. Combining centralized and P2P architectures, hybrid architectures allow users to interact through their mobile devices while accessing services from remote MSN servers.

Chang et al.² proposed a centralized architecture consisting of four main components: client devices, wireless-access network, the Internet and its hosts, and the server-side, including database- and application-specific servers. With a Web-service technology, the server can query a location module installed on the client device. The wireless-access network serves as TCP/IP pipes to allow the client and the server to communicate. The Internet component consists of third-party application servers (such as MapServer, the Simple Mail Transfer Protocol mail server, and Voice over IP).

Rana et al.¹⁹ proposed a service-oriented architecture with three main layers: service integrator, back-end services, and mobile client. The service integrator integrates mobile-device software and back-end services (such as for location tracking) through a standard interface. The back-end service layer is responsible for collecting Web data through special application programming interfaces (APIs) that set up connections between social networks and data-collector services. The service integrator ensures the interoperability of mobile clients with services. The mobile client allows client applications (such as Android and iOS) to access available services.

Johansson⁹ presented an architecture in which mobile devices (with audio, Bluetooth, and GPS) collect and process contextual information to assess its importance; data processing is handled by an MSN engine on the mobile device. The processed data is then used as input to the MSN.

Mani et al.¹² developed a software prototype that supports P2P spontaneous social networking through fast setup and deployment of a distributed social network that supports several services, including community creation, instant messaging, and VoIP.

Regardless of the type of architecture and number of layers or modules it includes, an MSN architecture must support several requirements: context awareness, acquiring and analyzing contextual data collected from Bluetooth, GPS, sensors, and other technologies; client/server and P2P communication, enabling mobile devices to communicate with each other, as well



Combining centralized and P2P architectures, hybrid architectures allow users to interact through their mobile devices while accessing services from remote MSN servers.



as with the server's back-end, to receive requested services; and services allowing generation of requested services and updating of related data. Hybrid architectures are best for addressing such requirements.

Classifying MSN Applications

We have identified four broad categories of MSN applications in the literature:

Pure and hybrid. Tong²⁴ classified MSN applications as pure and hybrid. Pure MSN applications are designed for mobile devices; hybrid MSN applications are designed for Web-based platforms but have been extended to mobile platforms;

Discovery. Pietiläinen¹⁸ proposed a categorization including three types of applications: proximity-based, check-in-based, and participatory sensing. Proximity-based MSN (such as Eagle,⁵ Kostakos and Neill,¹⁰ and Nicolai et al.¹⁶) uses devices that allow discovery of nearby devices. Related device information is useful for matching profiles on a central server,⁵ visualizing the surrounding social network on a mobile device,¹⁶ and displaying collected statistics of encounters with other users.¹⁰ Similar applications, including LoveGety and Nokia Sensor, have also been deployed in commercial settings that use Bluetooth to discover potential mates. In check-in-based MSN (such as BrightKite, Foursquare, Gowalla, Loopt, Mobiluck, and Whrrl), users constantly notify centralized Web servers of their current location and status. Mobile devices are just a way to update and consume available services. In participatory sensing,¹⁵ mobile devices collaboratively collect data from sensors (such as accelerometers, cameras, and GPS). Data is typically stored on central servers that provide aggregated reports of the data through a Web-based interface;

Major features. O'Sullivan¹⁷ proposed a classification system including six groups based on dominant features. In the texter group, the service focuses on sending short, text-based messages to a group of people simultaneously. In the radar group, the service knows the locations of users and their friends; applications allow users to check for nearby friends and/or receive notification if friends are nearby. In the geotagger group, MSN applica-

tions allow users to tag locations with images and information on a world map. Users may tag places for shopping, dining, or other activities, sharing the tags with friends. In the dating group, applications are identical to their online counterparts in which users create profiles for helping identify one another. In the social-networker group, applications aim to be like online social-networking platforms. And in the media-share group, applications share media files with groups of people; and

Push and pull. Rana et al.¹⁹ divided MSN applications into push-and-pull categories according to how data is acquired. Pull applications collect real-time information (such as micro-blogs and status) from various social networks using social APIs. Push applications are able to publish users' contextual information collected by sensors to various mobile networks.

Proposed classifications for MSN applications. Though Tong's proposed classification system²⁴ focuses on the nature of MSN applications by dividing them into pure and hybrid categories, the classification does not account for interaction between the application and the user (ultimately the mobile device), as highlighted in Pietiläinen's classification system.¹⁸ Neither classification emphasizes the services provided by MSN applications. The issue of services is the basis of the proposed grouping in O'Sullivan.¹⁷ This grouping does not seem accurate in light of the overlap between some groups that combine location-based services, messaging, media sharing, and geotagging. Rana et al.'s proposal¹⁹ is restrictive because it focuses on acquisition of data from user devices, as well from other social networks. Consequently, these proposals are inadequate for classifying MSN applications. We therefore propose to group MSN applications according to their categories, audience, usage, and interaction approaches (see Figure 2). The category group classifies these applications into pure and hybrid, as in Tong.²⁴ The audience group classifies MSN applications with respect to whether they accommodate individuals of all interests and backgrounds or have a niche focus, catering to specific groups of people. The usage



Current information-exchange models provide little protection for user privacy; for example, Facebook requires users allow access to their personal information and associate that information with their identities.



group classifies MSN applications according to their purpose, which could be informational (such as informing communities of news and promotions and addressing everyday problems), professional (such as job seeking), educational (such as collaboration with fellow students), dating, multimedia and content sharing, and social connections (such as being in touch with friends). The interaction-approach group includes the three subclasses proposed by Pietiläinen,¹⁸ proximity-based, check-in-based, and participatory sensing. Table 2 includes a partial list of MSN applications based on the classification in Figure 2.

Trends, Challenges, Opportunities

MSN users constantly search for ways to interact, engage, and share information while on the move through mobile devices (such as smartphones and tablets). Some newer devices support fourth-generation communication technologies, motivating vendors to provide services on a range of platforms, including Android, BlackBerry, iOS, and Windows 8. In addition to hardware improvement, application developers are moving toward mobile advertising, TV, and social gaming, as well as toward new services (such as mobile wallets), mobile commerce, and cloud-based services. These services are enticing research topics.

Emerging cloud-computing platforms (such as Amazon Web Services, Google App Engine, and Salesforce) can be coupled with mobile devices and MSN applications to create mobile social-cloud ecosystems in which MSN applications improve the user's experience and productivity, with cloud computing providing a robust, scalable, low-maintenance infrastructure. The mobile social cloud is driven in part by recent dramatic performance improvement in the IT infrastructure, together with innovations related to cloud computing (such as distributed computing, multicore processors, service-oriented architectures, and virtualization).

Meeting MSN-application-user expectations involves several challenges: One is performance, especially when users expect the same level of service on their mobile devices they enjoy on the desktop. Performance depends on available bandwidth of current cel-

lular networks, though it does not effectively support increased video-content exchange and delivery. For better performance, MSN applications must be able to cope with integration and standardization; social-networking stakeholders compete in the mobile-services market by proposing different proprietary solutions that do not find widespread acceptance due to integration difficulties. Manufacturers, designers, and developers must all agree on open solutions based on standards to address heterogeneity and interoperability of different hardware and software technologies.

Moreover, MSN application users meeting opportunistically through proximity-aware devices must be able to address challenges involved in maintaining efficient communications between mobile devices; for example, future wireless technologies (such as Bluetooth v3.0, low-power Wi-Fi, and Wi-Fi Direct) may someday support more-efficient opportunistic communications. Indeed, Bluetooth v3.0 includes native support for alternative physical layers to increase capacity while delivering low power consumption. Wi-Fi Direct promises to automate ad hoc device-to-device communications through 802.11.

Despite these advances, devel-

Table 2. Classifying MSN applications.

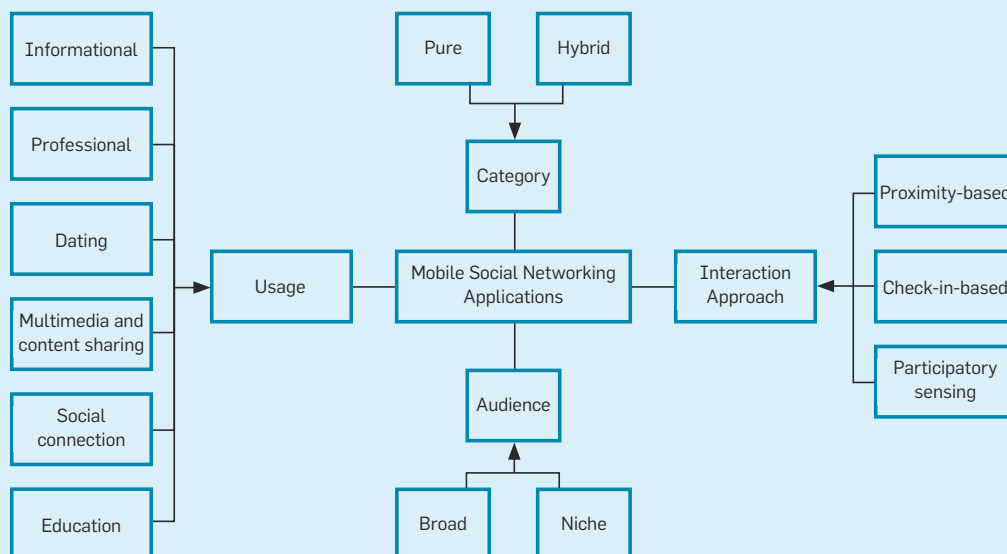
Class	Sub-Class	Examples
Category	Pure	Foursquare, Gowalla
	Hybrid	Facebook, MySpace, Twitter
Audience	Broad	Bebo, Facebook, MyYearBook, MySpace
	Niche	LoveGety, Orkut, PeopleNet, Speck
Features	Proximity-based	Jambo, FaceTime, LoveGety, Nokia Sensor, Proximating, Speck, Toothing ^{5,17}
	Check-in-based	Brightkite, Foursquare, Gowalla, Loopt; Mobiluck, Whrrl
	Participatory sensing	Wireless Rope ¹⁶
Use	Informational	PeopleNet, Serendipity
	Professional	LinkedIn
	Educational	iTeach
	Dating	Proximating, Toothing
	Social connections	Facebook, MySpace, Twitter
	Multimedia and content sharing	Friendster, YouTube, Zynga

opers must still guarantee efficient opportunistic communications, as new protocols and mechanisms are needed for the detection and control of temporal communities resulting from opportunistic communications. In addition to maintaining user privacy in these communities, researchers must begin to address how to aggregate the communities, looking for patterns in their creation and main-

tenance and improving content dissemination and resource sharing.

Another research challenge concerns design and implementation of adaptive discovery of friends or people sharing the same interests. Adaptation may help minimize the energy consumption of mobile devices, supporting dynamic changes in context and benefiting from historical information. Novel mechanisms may some

Figure 2. MSN applications classifications.



day support prediction of friends and identification of those nearby who share the same interests. Indeed, user A frequently detects friend B nearby but not friend C due to the limitations of proximity-aware devices. However, if friend B is able to detect friend C, then friend B is able to notify friend A that friend C is not far away and could be expected to appear soon. A can then be prepared to be in touch with C or alternatively might leave to avoid contact with C.

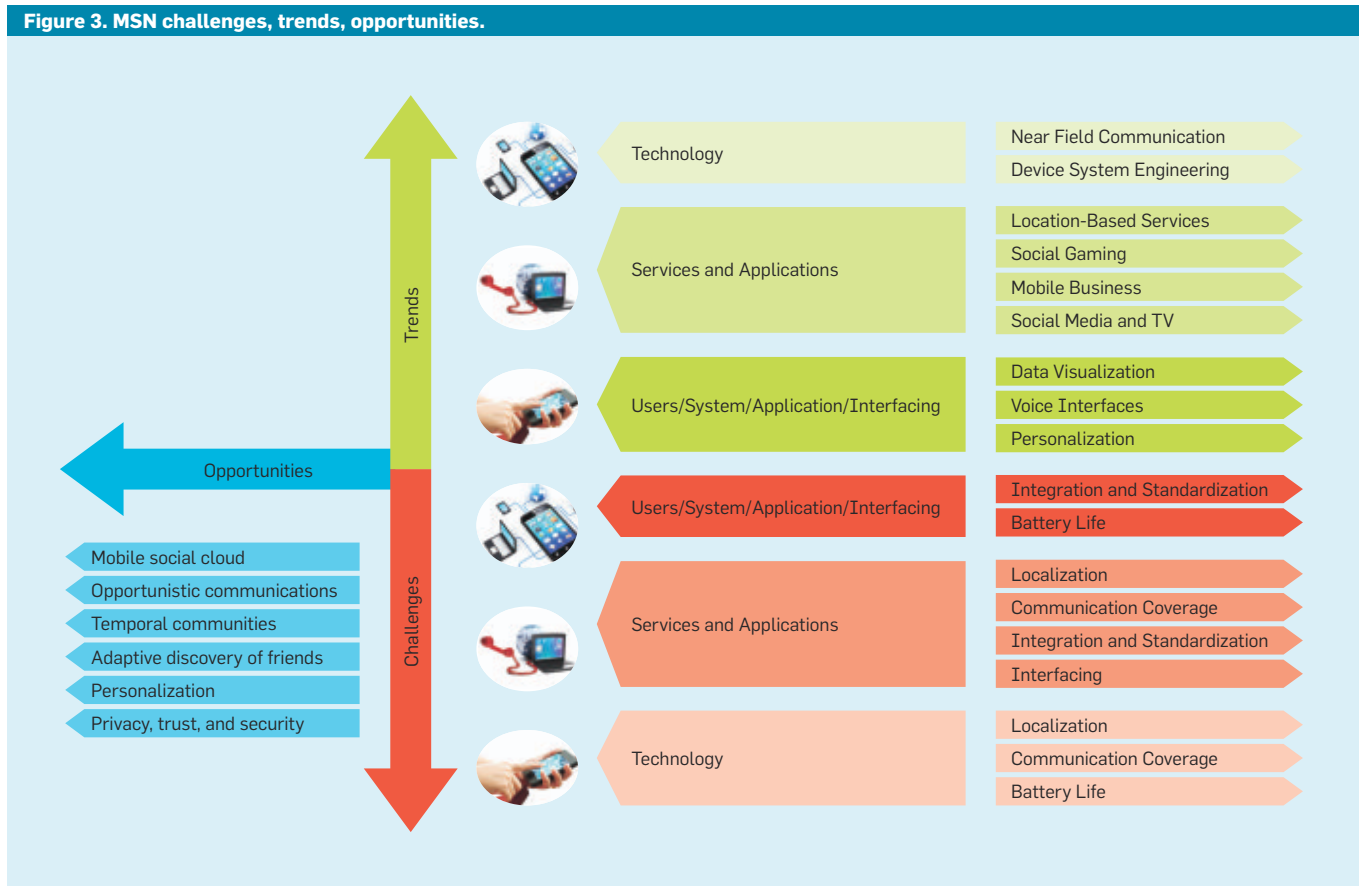
Personalization is another MSN challenge, with users requesting services with easy-to-use interfaces and the ability to match profiles, backgrounds, and contexts. It may be partially solved through regional languages in order to promote strong penetration of MSN applications; for example, Facebook recently launched a mobile application in multiple Indian regional languages, including Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, and Tamil. Since personalization requires knowledge of personal information, security and privacy is a serious challenge for developers of all MSN applications.

These applications must prevent data misuse or breaches of confidentiality, especially with businesses finding mobile phones to be at the core of their interest in collecting data and disseminating products. Current information-exchange models provide little protection for user privacy; for example, Facebook requires users allow access to their personal information and associate that information with their identities. In other systems using Bluetooth, nearby individuals can snoop other users' data sent openly through wireless connections. Access to user data makes it easy for malicious users to spoof and inject traffic into MSNs.¹ Knowing such attacks are possible, users lose trust in their service providers and fellow users, as they may not be the people they claim to be. In such instances, users are reluctant to disclose personal information, including identity and location. Using a trusted central server that collects information from individual users, computing and disseminating proximity results on demand, may guarantee privacy and security in MSN applications. However, such

a solution would be ineffective unless dedicated servers are deployed locally to support MSNs.⁴

Several recent projects offer solutions to privacy and security issues in MSN applications; for example, SocialAware¹ allows interaction of social-network information with real-world location-based services without compromising user privacy and security. Interaction is based on encrypted identifiers associated with a verified user location. The system then allows location-based services to query the local area for social-network information without disclosing personal user data.¹ Moreover, Dong et al.⁴ developed techniques and protocols for computing social proximity between any two users looking to discover potential friends. To prevent malicious hacker attacks (such as falsifying proximity) during exchange of attributes of the two users for whom the proximity is calculated, the authors developed a proximity-pre-filtering protocol to determine whether the proximity between users exceeds a given threshold. To protect privacy, the protocol ensures the initiator can

Figure 3. MSN challenges, trends, opportunities.



know only the comparison result between the estimated proximity and the threshold. To defend against attacks, Dong et al.⁴ developed a secure protocol based on homomorphic cryptography consisting of three major components: authentication without long-term linkability; efficient and privacy-preserving proximity pre-filtering; and private, verifiable proximity computation. Dong et al. also developed another protocol that leverages both homomorphic cryptography and obfuscation.

Although a number of initiatives have been proposed for user privacy, users and service providers alike need more efficient, scalable mechanisms for future MSN applications. Current policies in MSN applications should be personalized. Researchers may thus propose and implement new models for privacy and trust where user context, goals, profile, and cultural background are included. These models should also give users greater control over their personal data (see Figure 3). Ensuring privacy and security are related to services and applications, technology, and user/system/application interfaces, as an integral part of secure personal communication.

Conclusion

Mobile devices have spurred explosive growth and deployment of services delivered through MSN applications. In order to deliver the appropriate service to the right user at the right time, they exploit context awareness and emerging communication technologies. However, this goal is not yet completely achievable, so additional effort is needed to capture and analyze contextual features to allow smooth migration of services from PC-based environments to a mobile environments. We have surveyed current commercial MSN applications, analyzing many proposed MSN architectures. We also proposed a better way to classify MSN applications, summarizing related trends and challenges. A hybrid architecture involving P2P communications assisted by online servers is the most flexible. Providers are willing to offer the right services through the right technology and appropriate interface,

es, but constraints involving mobile devices and dynamic contexts must still be addressed.

Acknowledgments

We thank the anonymous reviewers for suggestions and comments that helped us improve the content, quality, and presentation of this article. We also thank Moshe Vardi for his kind encouragement, time, and support throughout its preparation. 

References

1. Beach, A., Gartrell, M., Ray, B., and Han, R. *Secure Socialware: A Security Framework for Mobile Social Networking Applications*. Technical Report CU-CS-1054 09. Department of Computer Science, University of Colorado, Boulder, CO, June 2009.
2. Chang, Y.J., Liu, H.H., Chou, L.D., Chen, Y.W., and Shin, H.Y. A general architecture of mobile social network services. In *Proceedings of the IEEE International Conference on Convergence Information Technology* (Gyeongju, South Korea, Nov. 21–23, 2007), 151–156.
3. ComScore. *Mobile Future in Focus*, 2012; http://www.comscore.com/Press_Events/Presentations_Whitepapers/2012/2012_Mobile_Future_in_Focus
4. Dong, W., Dave, V., Qui, L., and Zhang, Y. Secure friend discovery in mobile social networks. In *Proceedings of IEEE INFOCOM* (Shanghai, Apr. 10–15, 2011), 1647–1655.
5. Eagle, N. and Pentland, A. Social serendipity: Mobilizing social software. *IEEE Pervasive Computing* 4, 2 (Apr.–June 2005), 28–34.
6. Eagle, N. and Pentland, A. Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing* 10, 4 (May 2006), 255–268.
7. Gillin, P. Business anywhere anytime. *Mobile Social Networking: The New Ecosystem*. Computerworld Communications Brief, 2008; http://us.blackberry.com/business/leading/mobile_socialnetworking.pdf
8. GSMArena. *Mobile Phone Usage Report 2011: The Things You Do*; http://www.gsmarena.com/mobile_phone_usage_survey-review-592.php
9. Johansson, F. *Extending Mobile Social Software with Contextual Information*. Umeå University, Umeå, Sweden, 2008; <http://www.signar.se/blog/wp-content/extending-mobile-social-software-with-contextual-information.pdf>
10. Kostakos, V. and Neill, E.O. Cityware: Urban computing to bridge online and real-world social networks. In *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*. IGI Global, Queensland University of Technology, Australia, 2008, 195–204.
11. Liaquat, H., Chung, K.S.K., and Hasan, S.T. Exploring temporal communication through social networks. In *Proceedings of the 11th International Conference on Human-Computer Interaction* (Rio de Janeiro, Sept. 10–14). Springer-Verlag Berlin, Heidelberg, 2007, 19–30.
12. Mani, M., Nguyen, A.M., and Crespi, N. What's Up 2.0: P2P spontaneous social networking. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications* (Mar. 9–13, 2009), 1–2.
13. Marci, L. and Monné, V. A survey of mobile social networking. In *Proceedings of a Symposium on Internetworking TTK Technical Reports TTK-CSE-B5* (Helsinki University, Finland, Nov. 13, 2009); http://www.cse.tkk.fi/en/publications/B/5/papers/Vilalba_final.pdf
14. Marmasse, N., Schmandt, C., and Spectre, D. WatchMe: Communication and awareness between members of a closely knit group. In *Proceedings of the Sixth International Conference on Ubiquitous Computing* (Nottingham, England, Sept. 7–10, 2004), 214–231.
15. Miluzzo, E., Lane, N.D., Fodor, K., Peterson, R., Lu, H., Musolesi, M., Eisenman, S. B., Zheng, X., and Campbell, A.T. Sensing meets mobile social networks: The design, implementation, and evaluation of the CenceMe application. In *Proceedings of the Sixth ACM Conference on Embedded Network Sensor Systems* (Raleigh, NC, Nov. 4–7). ACM Press, New York, 2008, 337–350.
16. Nicolai, T., Yoneki, E., Behrens, N., and Kenn, H. Exploring social context with the Wireless Rope. In *Proceedings of the On The Move to Meaningful Internet Systems Workshops Lecture Notes in Computer Science Volume 4277* (Montpellier, France, Oct. 29–Nov. 3, 2006), 874–883.
17. O'Sullivan, C. Mobile social networking. *GoMo News* (Nov. 4, 2012); <http://www.gomonevents.com/moso/>
18. Pietiläinen, A.K. *Opportunistic Mobile Social Networks at Work*. Ph.D. Thesis, Université Pierre et Marie Curie, Paris, 2010; <http://www.thlab.net/~apietila/pubs/thesis.pdf>
19. Rana, J., Kristiansson, J., Hallberg, J., and Synnes, K. Challenges for mobile social networking applications. In *Proceedings of Communications Infrastructure, Systems and Applications in Europe Lecture Notes of the Institute for Computer Sciences Volume 16* (London, Aug. 11–13, 2009), 241–246.
20. Siewiorek, D., Smailagic, A., Furukawa, J., Krause, A., Moraveji, N., Reiger, K., Shaffer, J., and Wong, F. SenSay: A context-aware mobile phone. In *Proceedings of the Seventh IEEE International Symposium on Wearable Computers* (Washington, D.C., Oct. 21–23, 2003), 248–249.
21. Smith, I. Social-mobile applications Reno. *Computer* 38, 4 (Apr. 2005), 84–85.
22. Tamarit, P., Calafate, C.T., Cano, J.C., and Manzoni, P. BlueFriend: Using Bluetooth technology for mobile social networking. In *Proceedings of the Sixth Annual International Conference on Mobile and Ubiquitous Systems* (Toronto, July 13–16). IEEE, 2009, 1–2.
23. Tode, Ch. Facebook rapidly turning into mobile powerhouse. *Mobile Marketer* (Dec. 30, 2011); <http://www.mobilemarketer.com/cms/news/social-networks/11804.html>
24. Tong, C. Analysis of some popular mobile social network systems. In *Proceedings of the Seminar on Internetworking* (Helsinki University, Finland, Apr. 28, 2008); http://www.cse.tkk.fi/en/publications/B/1/papers/Chang_final.pdf
25. Warren, C. Mobile social networking usage soars [stats]. *Mashable* (Mar. 3, 2010); <http://mashable.com/2010/03/03/comscore-mobile-stats/>

Nafaâ Jabeur (nafaâ@aou.edu.om) is an assistant professor in the Department of Information Technology and Computing at the Arab Open University, Oman Branch, Sultanate of Oman.

Sherali Zeadally (szeadally@udc.edu) is an associate professor in the Department of Computer Science and Information Technology at the University of the District of Columbia, Washington, D.C.

Biju Sayed (b_sayed@du.edu.om) is an assistant professor in the Department of Computer Science at Dhofar University, Oman.

**Discovering surprises in
the face of intractability.**

BY FEDOR V. FOMIN AND PETTERI KASKI

Exact Exponential Algorithms

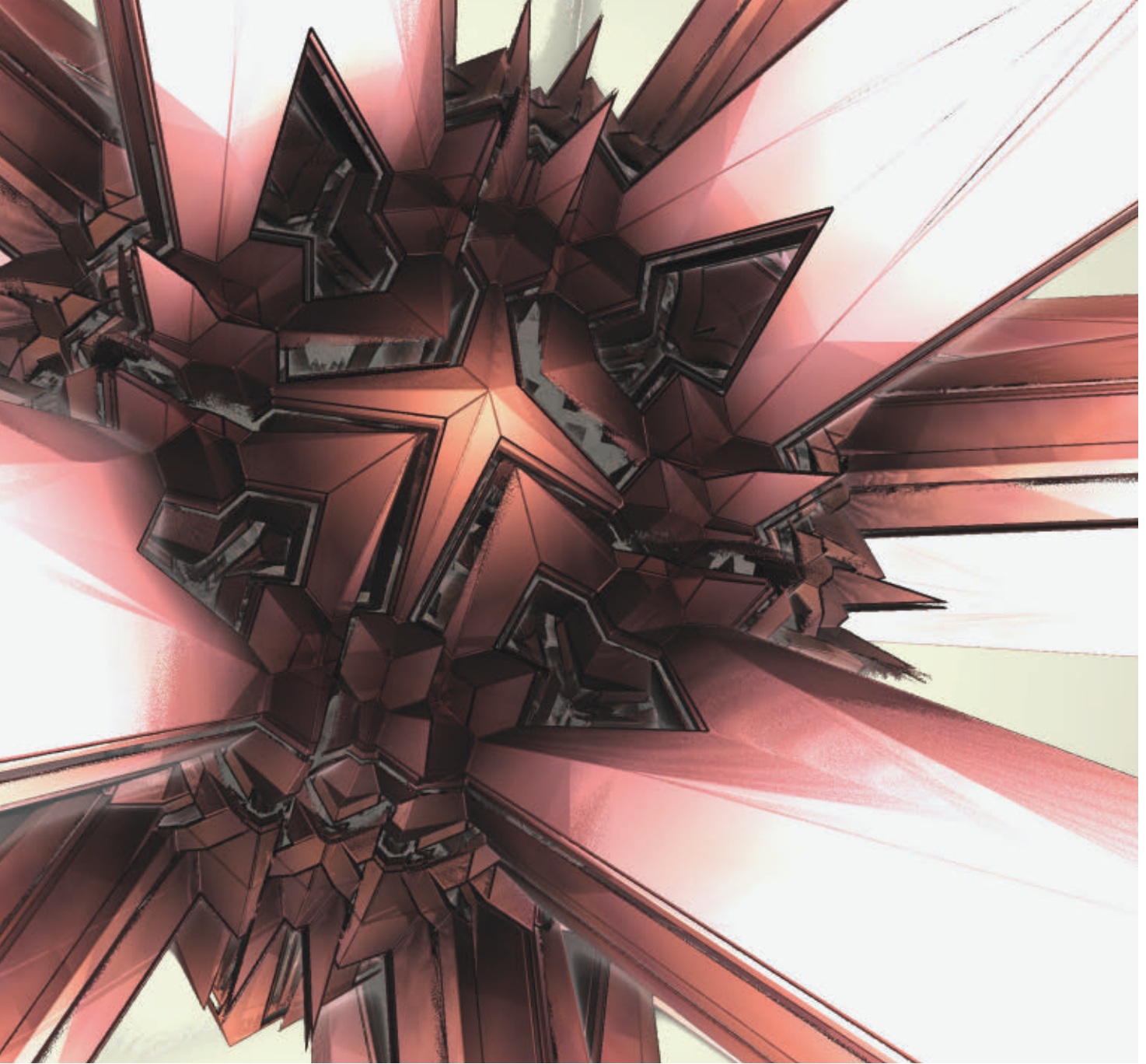
MANY COMPUTATIONAL PROBLEMS have been shown to be intractable, either in the strong sense that no algorithm exists at all—the canonical example being the undecidability of the Halting Problem—or that no *efficient* algorithm exists. From a theoretical perspective perhaps the most intriguing case occurs with the family of *NP*-complete problems, for which *it is not known* whether the problems are intractable. That is, despite extensive research, neither is an efficient algorithm known, nor has the existence of one been rigorously ruled out.¹⁶

To cope with intractability, advanced techniques such as *parameterized algorithms*^{10,13,31} (that isolate the exponential complexity to a specific structural parameter of a problem instance) and *approximation algorithms*³⁴ (that produce a solution whose value is guaranteed to be within a known factor of the value of an optimum solution) have been developed. But what can we say about finding exact solutions

of non-parameterized instances of intractable problems? At first glance, the general case of an *NP*-complete problem is a formidable opponent: when faced with a problem whose instances

» key insights

- While it remains open whether or not *P* equals *NP*, significant progress in the area of exhaustive search has been made in the last few years. In particular, many *NP*-complete problems can now be solved significantly faster by exhaustive search. The area of exact exponential algorithms studies the design of such techniques.
- While many exact exponential algorithms date back to the early days of computing, a number of beautiful surprises have emerged recently.



can express arbitrary nondeterministic computation, how is one to proceed at solving a given instance, apart from the obvious exhaustive search that “tries out all the possibilities”?

Fortunately, the study of algorithms knows many positive surprises. Computation is malleable in nontrivial ways, and subtle algorithmically exploitable structure has been discovered where none was thought to exist. Furthermore, the more generous a time budget the algorithm designer has, the more techniques become available. Especially so if the budget is exponential in the size of the input. Thus, absent complexity-theoretic obstacles, *one should be able to do better than exhaustive*

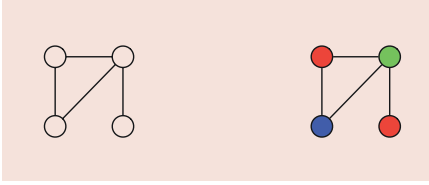
search. This is the objective of *exact exponential algorithms*.¹⁵

Arguably, the oldest design technique to improve upon exhaustive search is *branching* or *backtrack search*,^{18,35} which recursively splits the exhaustive search space, attempting to infer in the process that parts of the space need not be visited. For recent applications of branching techniques, we refer to Eppstein¹² and Fomin et al.¹⁴ Another classical design technique is *dynamic programming*,² which derives a solution from the bottom up by storing solutions of smaller subproblems and combining them via a recurrence relation to progressively yield solutions of larger subproblems. These two techniques in many cases give significant

improvements over plain exhaustive search, but in other cases, no improvement at all upon exhaustive search has been available, and many problems remain with this status.

In what follows, we do not try to give a comprehensive survey of exact exponential algorithms. Indeed, even listing the most significant results would require a format different from this review. Instead, we have chosen to review the area by highlighting three recent results. In each case, research had been essentially stuck for an extended period of time—in one case for almost 50 years!—and it was conceivable that perhaps no improvement could be obtained over the known algorithms. But computation has the power

Figure 1. Graph coloring.



to surprise, and in this article we hope to convey some of the excitement surrounding each result. We also find these results particularly appealing because they are *a posteriori* quite accessible compared with many of the deep results in theoretical computer science, and yet they illustrate the subtle ways in which computation can be orchestrated to solve a problem.

Three NP-Complete Problems

The three problems we discuss in more detail are *Maximum 2-Satisfiability*, *Graph Coloring*, and *Hamiltonian Path*. We start by giving an overview of previous approaches to attack each problem, and then in the subsequent sections discuss the novel algorithms.

MAX-2-SAT. The satisfiability problem takes as input a logical expression built from n variables $x_1, x_2, \dots, x_n$ and the Boolean connectives $\neg$ (NOT), $\vee$ (OR), and $\wedge$ (AND). The task is to decide whether the expression can be *satisfied* by assigning a truth value, either 0 (false) or 1 (true), to each variable such that the expression evaluates to 1. For example, the expression

$$(x_1 \vee \neg x_2) \wedge (\neg x_1 \vee \neg x_2) \wedge (x_1 \vee x_2) \quad (1)$$

can be satisfied by setting $x_1 = 1$ and $x_2 = 0$, whereas the expression

$$\begin{aligned} &(x_1 \vee \neg x_2 \vee x_3) \wedge (\neg x_1 \vee x_2 \vee \neg x_3) \\ &\wedge (\neg x_1 \vee \neg x_2 \vee x_3) \wedge (x_1 \vee x_2 \vee x_3) \\ &\wedge (x_1 \vee x_2 \vee \neg x_3) \wedge (x_1 \vee \neg x_2 \vee \neg x_3) \\ &\wedge (\neg x_1 \vee \neg x_2 \vee \neg x_3) \wedge (\neg x_1 \vee x_2 \vee x_3) \end{aligned} \quad (2)$$

is not satisfiable.

It is customary to assume that the input expression is in *conjunctive normal form*, where it is required that the expression is the AND of *clauses*, each of which is an OR of *literals*, which are variables or negations of variables. If all clauses have k literals, then the expression is in *k-conjunctive normal form*, or *k-CNF*. For example, (1) is in 2-CNF and (2) is in 3-CNF. The satisfiability

problem for an expression in *k-CNF* is called the *k-CNF satisfiability* or *k-SAT* problem. It is polynomial-time solvable for $k \leq 2$ and NP-complete for $k \geq 3$.¹⁷

A stronger variant of the problem, *maximum k-CNF satisfiability* or *MAX-k-SAT*, gives a threshold t as additional input, and the task is to decide whether there is an assignment of truth values to the variables such that at least t clauses evaluate to 1. This variant is NP-complete for all $k \geq 2$.¹⁷

MAX-k-SAT is trivially solvable by trying all possible truth assignments. When a formula has n variables, it has 2^n possible assignments and for each assignment we can compute in polynomial time how many clauses are satisfied. Thus, the total running time, up to a factor polynomial in n , is dominated by 2^n . A special case of the problem, known as MAX-CUT, can be obtained by formulating MAX-2-SAT as a problem of partitioning the vertices of an n -vertex graph into two subsets such that at least t edges cross between subsets. However, even in the special cases of MAX-2-SAT and MAX-CUT, no better algorithm than the trivial exhaustive search was known until the work of Williams.³⁶

Graph Coloring. In the graph coloring problem, we are given as input a graph G with n vertices and a palette of k colors. The task is to decide whether it is possible to assign to each vertex a color from the palette so that the coloring is *proper*, that is, every edge has distinct colors at its ends. For example, the graph in Figure 1 admits a proper coloring of its vertices using three colors.

The graph coloring problem is polynomial-time solvable for $k \leq 2$ and NP-complete for $k \geq 3$.¹⁷ The minimum number of colors for which a graph G has a proper coloring is the *chromatic number* $\chi(G)$ of G .

The first algorithmic approaches to compute the chromatic number of a graph can be traced back to the work of Zykov.⁴¹ The idea is based on a branching procedure. The base case of the branching occurs when all pairs of vertices of G are adjacent, that is, G is a complete graph, in which case the chromatic number is equal to the number of vertices in G . Otherwise, G contains a pair u, v of vertices that are not joined by an edge. In every proper coloring of G it holds that u and v either

have distinct colors (in which case we construct a new graph by joining u and v with an edge), or have the same color (in which case we construct a new graph by identifying u and v). This enables us to recursively branch on the two cases and return the best of the two solutions obtained. In terms of running time, however, this approach is in general no better than plain exhaustive search, which involves iterating through the k^n distinct ways to color the n vertices of G using the k available colors, and for each coloring testing whether it is proper.

After Zykov's seminal work, the history of algorithms for graph coloring benefits from a digression to the study of independent sets in graphs. In particular, every proper coloring of G has the property that no two vertices of the same color are joined by an edge. Such a set of vertices is an *independent set* of G . An independent set of G is *maximal* if it is not a proper subset of a larger independent set of G . In 1976, Lawler²⁷ observed that dynamic programming and advances in the study of independent sets can be used to drastically improve upon the k^n exhaustive search. Let us first develop a basic version of the algorithm. Since each color class in a proper coloring of G is an independent set of G , we have that G is k -colorable if and only if the vertex set V of G decomposes into a union of k independent sets of G . Stated in terms of the chromatic number, we have $\chi(G) = 0$ if G has no vertices; otherwise, we have

$$\chi(G) = 1 + \min \{ \chi(G \setminus I) : I \in \mathcal{I}(G) \}, \quad (3)$$

where $\mathcal{I}(G)$ is the family of all nonempty independent sets of G , and $G \setminus I$ denotes the graph obtained from G by deleting the vertices in I . For every subset $X \subseteq V$, we can thus compute the chromatic number $\chi(G[X])$ of the subgraph of G induced by X as follows. When X is empty, we set $\chi(G[X]) = 0$. When X is nonempty, we compute the value $\chi(G[X])$ from the already computed values of proper subsets of X by making use of (3).

What is the running time of this algorithm? The algorithm considers all subsets $X \subseteq V$, and for each such X , it considers all $I \subseteq X$ that are independent in $G[X]$. The number of such I is at most $2^{|X|}$. Thus, the number of steps of

the algorithm is, up to a factor polynomial in n , at most $\sum_{i=0}^n \binom{n}{i} 2^i = 3^n$.

Lawler also observed that the basic 3^n -algorithm can be improved. Namely, instead of going through all subsets $I \subseteq X$ that are independent in $G[X]$, it suffices to consider only maximal independent sets of $G[X]$. It was known²⁹ already in the 1960s that the number of maximal independent sets in a graph with i vertices is at most $3^{i/3}$, and that these sets can be listed in time $\mathcal{O}(3^{i/3}n)$. Thus, the exponential part of the running time of the algorithm is bounded by

$$\sum_{i=0}^n \binom{n}{i} 3^{i/3} = (1 + \sqrt[3]{3})^n < 2.45^n.$$

It is possible to make even further improvements of this idea by more accurate counting of large and small maximal independent sets.¹¹ But in all these improvements the following common pattern seemed unavoidable: we have to go through all vertex subsets of the graph, and for each subset, we have to enumerate an exponential number of subsets, resulting in time C^n , for a constant $C > 2$.

Hamiltonian Path. In the NP-complete *Hamiltonian cycle* problem, we are given a graph on n vertices and the task is to decide whether the graph has a Hamiltonian cycle, which is a cycle visiting every vertex of the graph exactly once. For example, the graph in Figure 2 has a Hamiltonian cycle, outlined in bold edges.

This is a special case of the famous *Traveling Salesman Problem*, where the task is to, given an $n \times n$ matrix of travel costs between n cities, design a travel schedule that visits each city exactly once and returns back to the starting point so that the total cost is minimized.

A stronger variant, the *Hamiltonian path* problem, constrains one of the vertices as the first vertex s and another vertex t as the last vertex, and asks us to decide whether the graph has a path that starts at s , ends at t , and visits all the vertices exactly once. (By trying all the pairs $\{s, t\}$ joined by an edge, we can solve the Hamiltonian cycle problem if we can solve the Hamiltonian path problem.)

For the Hamiltonian path problem, exhaustive search iterates through the $(n-2)!$ ways to arrange the n vertices into a sequence that starts at s and ends at t , testing for each sequence whether it forms a path (of the minimum cost).

Bellman³ and Held and Karp¹⁹ used dynamic programming to solve the problem in time $\mathcal{O}(2^n n^2)$, by keeping track for every vertex v and vertex subset S , the existence (or the minimum cost) of a path from s to v that visits exactly the vertices in $S \subseteq V$. This algorithm, however, requires also space 2^n .

It is possible to solve the problem within the same running time but within polynomial space by making use of the principle of inclusion and exclusion. It seems that essentially the same approach was rediscovered several times.^{1,23,25} To illustrate the design, Figure 3 displays a graph with $n = 8$ vertices $\{a, b, c, d, e, f, g, h\}$.

Let us assume that $s = a$ and $t = h$. A walk of length $n-1$ that starts from s and ends at t can be viewed as a string of length $2n-1$ with alternating and possibly repeating vertices and edges, such as

$$aAeCbDfFcGgIdJh \quad (4)$$

or

$$aBfDbEgGcFfFcHh. \quad (5)$$

We observe that each such walk makes exactly n visits to vertices and contains, possibly with repetitions, $n-1$ edges. Moreover, the walk is a Hamiltonian path if and only if the walk visits n distinct vertices; indeed, otherwise there is at least one vertex that is visited more than once. For example, (4) is a path and (5) is a non-path because it repeatedly visits f (and c).

Although finding a Hamiltonian path is a challenging computational problem, one can compute in polynomial time the number of walks of length k from s to t . Indeed, let A be the adjacency matrix of G with rows and columns indexed by vertices of G , such that the (x, y) -entry of A is set to 1 if there is an edge from x to y in G , and set to 0 otherwise. By induction on k we observe that the (s, t) -entry of the k th matrix power A^k counts the number of walks of length k in G that start at s and end at t . Therefore, the number of walks of length $n-1$ can be read from the matrix A^{n-1} , which can be computed in time polynomial in n .

One approach to isolate the paths among the walks is to employ the *principle of inclusion and exclusion*. Consider a finite set X and three subsets A_1, A_2 , and A_3 (see Figure 4).

To obtain $|A_1 \cup A_2 \cup A_3|$, we can use the following formula

Figure 3. Example for Hamiltonian path.

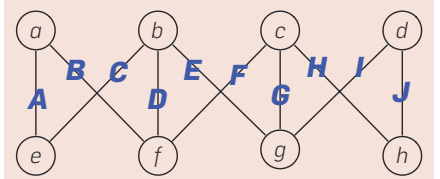


Figure 2. Hamiltonian cycle.

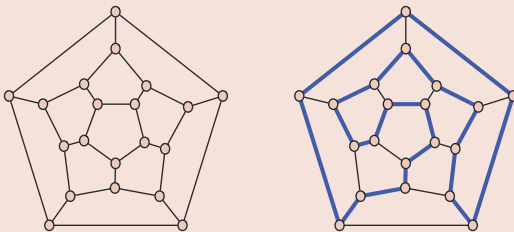


Figure 4. A Venn diagram for three subsets.

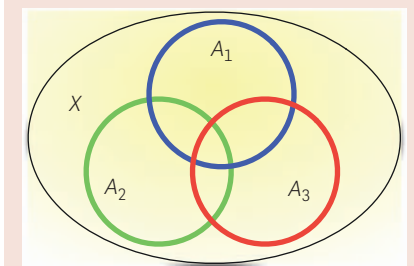
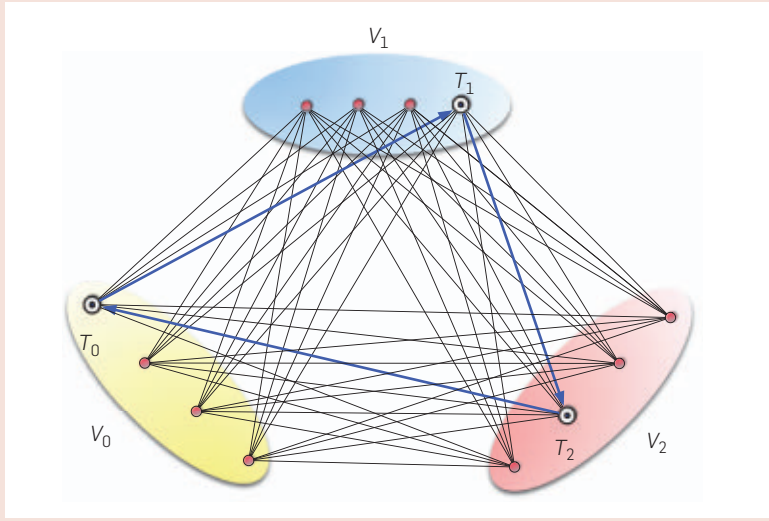


Figure 5. The directed graph D with one triangle $T_0T_1T_2$ highlighted.

$$|A_1 \cup A_2 \cup A_3| = |A_1| + |A_2| + |A_3| \\ - |A_1 \cap A_2| - |A_1 \cap A_3| \\ - |A_2 \cap A_3| + |A_1 \cap A_2 \cap A_3|,$$

or, equivalently,

$$|X \setminus (A_1 \cup A_2 \cup A_3)| = |X| - |A_1| - |A_2| - |A_3| \\ + |A_1 \cap A_2| + |A_1 \cap A_3| \\ + |A_2 \cap A_3| \\ - |A_1 \cap A_2 \cap A_3|.$$

The principle of inclusion and exclusion generalizes the last formula to the case when there are q subsets $A_1, A_2, \dots, A_q$ of X by

$$\left| X \setminus \bigcup_{i=1}^q A_i \right| = \sum_{J \subseteq \{1, 2, \dots, q\}} (-1)^{|J|} \left| \bigcap_{j \in J} A_j \right|. \quad (6)$$

Let us come back to the Hamiltonian path problem. Take $q = n - 2$ and suppose that the vertices other than s and t are labeled with integers $1, 2, \dots, n - 2$. Let X be the set of all walks of length $n - 1$ from s to t and, for each $i = 1, 2, \dots, n - 2$, let A_i be the set of walks in X that avoid the vertex i . Then, $X \setminus \bigcup_{i=1}^q A_i$ is the set of Hamiltonian paths, and we can use (6) to count their number. In particular, for each fixed $J \subseteq \{1, 2, \dots, q\}$, the right-hand side of (6) can be computed time polynomial in n by counting the number of walks of length $n - 1$ from s to t in the graph with the vertices in J deleted.

This approach can be used to

compute the number of Hamiltonian paths in an n -vertex graph in time $\mathcal{O}(2^n n)$. It is also possible to obtain similar running time by making use of dynamic programming. But in both approaches, it seemed that the most time consuming part of the procedure, going through all possible vertex subsets, was unavoidable. This situation was particularly frustrating because the 2^n barrier had withstood attacks since the early 1960s.

Surprise 1: MAX-2-SAT

Let us recall that for MAX-2-SAT the challenge was to break the 2^n barrier in running time. The following approach for doing this is due to Williams.³⁶ An alternative approach via sum-product algorithms is due to Koivisto.²⁶

Let us start with a seemingly unrelated task, namely that of deciding whether a given directed graph D contains a *triangle*, that is, a triple x, y, z of vertices such that the arcs xy, yz , and zx occur in D . While the immediate combinatorial approach to find a triangle in a v -vertex graph is to try all possible triples of vertices, which would require $\mathcal{O}(v^3)$ steps, there is a faster algorithm of Itai and Rodeh.²² The algorithm relies on formulating the problem in terms of linear algebra. Let A be the adjacency matrix of D , and recall that the (s, t) -entry of the k th power A^k counts the number of walks of length k from s to t . In particular, every walk of

length 3 that starts and ends at a vertex x must pass through three distinct vertices, and thus form a triangle, enabling us to extract the number of triangles in D from the diagonal entries of the matrix A^3 . Thus, it suffices to compute the matrix A^3 . The immediate algorithm for computing the product of two $v \times v$ matrices requires $\mathcal{O}(v^3)$ steps. However, this product can be computed in time $\mathcal{O}(v^\omega)$, where $\omega < 2.376$ is the so-called square matrix multiplication exponent; see Coppersmith and Winograd⁷ and Strassen.³² Very recently, it has been shown that $\omega < 2.3727$.³³

The key insight is now to exploit the fact that triangles can be found quickly to arrive at a nontrivial algorithm for MAX-2-SAT. Toward this end, suppose we are given as input a 2-CNF formula F over n variables. We may assume that n is divisible by 3 by inserting dummy variables as necessary. Let X be the set of variables of F and let X_0, X_1, X_2 be an arbitrary partition of X into sets of size $n/3$.

Let us transform the instance F into a directed graph D as follows. For every $i = 0, 1, 2$ and every subset $T_i \subseteq X_i$, the graph D has a vertex T_i . The meaning of T_i is that it corresponds to an assignment that sets all variables in T_i to the value 1 and all variables in $X_i \setminus T_i$ to the value 0. Let us write V_i for the set of all subsets $T_i \subseteq X_i$. The arcs of D are all possible pairs of the form (T_i, T_j) , where $T_i \subseteq X_i, T_j \subseteq X_j$, and $j \equiv i + 1 \pmod{3}$. We observe that D has $v = 3 \times 2^{n/3}$ vertices and $3 \times 2^{2n/3}$ arcs. For $i = 0, 1, 2$, let the set C_i consist of all clauses of F that either (a) contain variables only from X_i ; or (b) contain one variable from X_i and one variable from X_j , with $j \equiv i + 1 \pmod{3}$. Now observe that every clause of F has at most two variables. In particular, either both these variables belong to some set X_i , or one variable is in X_i and the other is in X_j with $j \equiv i + 1 \pmod{3}$. Thus, the sets C_0, C_1, C_2 partition the clauses in F . We still require weights on the arcs of D . Let us set the weight $w(T_i, T_j)$ of the arc from $T_i \subseteq X_i$ to $T_j \subseteq X_j$ to be equal to the number of clauses in C_i satisfied by assigning the value 1 to all variables in $T_i \cup T_j$ and the value 0 to all remaining variables in $(X_i \cup X_j) \setminus (T_i \cup T_j)$.

To illustrate the construction, let us assume F is the following formula

$$\begin{aligned}
& (x_1 \vee x_2) \wedge (\neg x_2 \vee x_3) \wedge (x_1 \vee x_3) \\
& \wedge (\neg x_2 \vee x_4) \wedge (x_3 \vee x_4) \\
& \wedge (x_1 \vee \neg x_5) \wedge (\neg x_4 \vee \neg x_6)
\end{aligned}$$

and partition the variables so that $X_0 = \{x_1, x_2\}$, $X_1 = \{x_3, x_4\}$, and $X_2 = \{x_5, x_6\}$. Then, $C_0 = \{(x_1 \vee x_2), (\neg x_2 \vee x_3), (x_1 \vee x_3), (\neg x_2 \vee x_4)\}$, $C_1 = \{(x_3 \vee x_4), (\neg x_4 \vee \neg x_6)\}$, and $C_2 = \{(x_1 \vee \neg x_5)\}$. Figure 5 illustrates the underlying graph D , where each set V_0, V_1, V_2 has size 4. For example, $V_0 = \{\emptyset, \{x_1\}, \{x_2\}, \{x_1, x_2\}\}$. For sets $T_0 = \emptyset$, $T_1 = \{x_3, x_4\}$, and $T_2 = \{x_6\}$, the corresponding assignment, viz. $x_1 = x_2 = 0$, $x_3 = 1$, $x_4 = 1$, $x_5 = 0$, $x_6 = 1$, satisfies five clauses. Accordingly, the weight of the triangle $T_0 T_1 T_2$ in D is also five.

The equivalence of the following statements follows from the construction of D : (i) There is a subset of variables $T \subseteq X$ such that exactly t clauses are satisfied by assigning the value 1 to variables in T and the value 0 to the variables in $X \setminus T$. (ii) The graph D contains a triangle $T_0 T_1 T_2$ with $T_i \subseteq X_i$ for each $i = 0, 1, 2$ such that

$$t = w(T_0, T_1) + w(T_1, T_2) + w(T_2, T_0).$$

Thus, to find an assignment that satisfies most clauses, it suffices to find a heaviest triangle in D .

We are almost done. Indeed, every formula with n variables has at most $4n^2$ clauses of length 2, and hence to find a heaviest triangle, it suffices to test for the existence of a triangle of weight t for each $0 \leq t \leq 4n^2$ in turn. To test for a triangle of weight t , we go through all possible $\mathcal{O}(t^3)$ partitions $t = t_0 + t_1 + t_2$ into nonnegative parts, and for each partition, we construct a subgraph D_{t_0, t_1, t_2} of D by leaving only arcs of weight t_i for arcs going from subsets of X_i to subsets of X_j with $j \equiv i + 1 \pmod{3}$. Finally, it suffices to decide whether D_{t_0, t_1, t_2} has a triangle. The subgraph D_{t_0, t_1, t_2} can be constructed in time $\mathcal{O}(2^{2n/3}n)$ by going through all arcs of D . The total running time is thus

$$\mathcal{O}\left(\sum_{t=0}^{4n^2} t^3 (2^{n/3} + 2^{2n/3})\right).$$

Because $\omega < 2.376$, we conclude that the running time of the algorithm is $\mathcal{O}(1.74^n)$.

Surprise 2: Graph Coloring

The next surprise is due to Björklund et al.⁶ To explain the idea of the algorithm, it will again be convenient to start with a task that may appear at first completely unrelated, namely the multiplication of polynomials. To multiply two given polynomials, the elementary algorithm is to cross-multiply the monomials pairwise and then collect to obtain the result:

$$\begin{aligned}
(1 + 3x + x^2)(2 - x + x^2) \\
= 2 - x + x^2 + 6x - 3x^2 + 3x^3 + 2x^2 \\
- x^3 + x^4 \\
= 2 + 5x + 2x^3 + x^4.
\end{aligned}$$

In particular, if we are multiplying two polynomials of degree d (that is, the highest degree of a monomial with a nonzero coefficient is d), we require $\mathcal{O}(d^2)$ steps to get the result via the elementary algorithm due to the cross-multiplication of monomials. Fortunately, we can drastically improve upon the elementary algorithm by deploying the fast Fourier transform (FFT) to evaluate both input polynomials (given as two lists of $d + 1$ coefficients, one coefficient for each monomial) at $2d + 1$ distinct points, $x_0, x_1, \dots, x_{2d}$, then multiplying the evaluations pointwise, and finally employing the inverse FFT to recover the list of coefficients for the product polynomial. With such an algorithm, the number of operations is reduced from $\mathcal{O}(d^2)$ to $\mathcal{O}(d \log d)$.

But what about graph coloring? Could we formulate the task of decomposing the vertex set into a union of independent sets of G as a task *analogous* to polynomial multiplication? Let us try to find the solution incrementally for $j = 1, 2, \dots, k$. Suppose we have a list of all the sets of vertices that decompose into a union of j independent sets of G , and would like to determine such a list for $j + 1$.

Let us consider an example. Figure 6

displays a graph with $n = 4$ whose independent sets are

$$\emptyset, \{a\}, \{b\}, \{c\}, \{d\}, \{a, d\}, \{c, d\}.$$

For $j = 2$, the sets of vertices that decompose into a union of j independent sets are

$$\begin{aligned}
& \emptyset, \{a\}, \{b\}, \{c\}, \{d\}, \{a, b\}, \{a, c\}, \\
& \{a, d\}, \{b, c\}, \{b, d\}, \{c, d\}, \\
& \{a, b, d\}, \{a, c, d\}, \{b, c, d\}.
\end{aligned}$$

Given the family of independent sets and the family of solutions for j , we would like to determine the family of solution for $j + 1$. Pursuing an analogy with polynomial multiplication, we can view the sets in both set families as “monomials” and multiply these “monomials” using set union. For example:

$$\begin{aligned}
& (\emptyset + \{a\} + \{a, b\}) \cup (\emptyset + \{b, c\} + \{c\}) \\
& = \emptyset + \{b, c\} + \{c\} \\
& \quad + \{a\} + \{a, b, c\} + \{a, c\} \\
& \quad + \{a, b\} + \{a, b, c\} + \{a, b, c\} \\
& = \emptyset + \{a\} + \{a, b\} + \{a, c\} \\
& \quad + 3\{a, b, c\} + \{b, c\} + \{c\}.
\end{aligned}$$

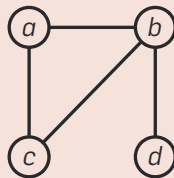
In general, both set families being multiplied may have up to 2^n members, and the same holds for the product. Again the elementary algorithm will consider the monomials pairwise, which requires consideration of $2^n \times 2^n = 4^n$ pairs in the worst case. But analogous to polynomial multiplication, it turns out that we can do considerably better.

Suppose the input set families are f and g . We can view f (and similarly g) as a function that takes an integer value $f(S)$ for each subset $S \subseteq V$ of our n -element vertex set V . (Indeed, let us assume that we have $f(S) = 1$ if and only if the set S is in the family, and $f(S) = 0$ otherwise.) The product, $e = f \cup g$, is then a similar function defined for each $S \subseteq V$ by the rule

$$e(S) = \sum_{A, B \subseteq V: A \cup B = S} f(A)g(B).$$

Since each pair (A, B) contributes by $f(A)g(B)$ to the value of e at exactly $S = A \cup B$, we observe that $\mathcal{O}(4^n)$ multiplications and additions suffice to compute the function e from the given functions f and g , which corresponds to the elementary multiplication algorithm. Now, the analogy to the FFT algorithm

Figure 6. Example for graph coloring.



for multiplying polynomials suggests a different approach, namely to transform the inputs f and g somehow, then multiply pointwise, and finally transform back to the original representation to recover $f \cup g$. The relevant transform turns out to be the *zeta transform* $f\zeta$ of f , defined for all $Y \subseteq V$ by

$$f\zeta(Y) = \sum_{X \subseteq Y} f(X),$$

and its inverse, the *Möbius transform* $f\mu$ of f , defined for all $Y \subseteq V$ by

$$f\mu(Y) = (-1)^{|Y|} \sum_{X \subseteq Y} (-1)^{|X|} f(X).$$

Indeed, the product $f \cup g$ can be computed using the expression

$$f \cup g = ((f\zeta) \times (g\zeta))\mu.$$

Both the zeta transform $f \mapsto f\zeta$ and the Möbius transform $f \mapsto f\mu$ admit fast algorithms analogous to the FFT. Indeed, it follows from the work of Yates⁴⁰ (see Knuth²⁴) that given f as input, we can compute $f\zeta$ (and similarly $f\mu$) using $\mathcal{O}(2^n n)$ additions and subtractions. This algorithm is perhaps best illustrated in arithmetic circuit form, which Figure 7 illustrates in the case $n = 3$. Observe that each of the n dashed cubes takes the sum along one of the n “dimensions” so that each output $f\zeta(Y)$ ends up taking the sum of all the inputs $f(X)$ with $X \subseteq Y$.

We can thus compute $e = f \cup g$ from f and g given as input using $\mathcal{O}(2^n n)$ additions, negations, and multiplications.

It now follows that we can decide in $\mathcal{O}(2^n nk)$ steps whether a given n -vertex graph G is k -colorable. Indeed, we first compute the characteristic function f of the independent sets of G , that is, for each $S \subseteq V$ we set $f(S) = 1$ if S is independent in G , and $f(S) = 0$ otherwise.

Next, we compute the functions e_j for $j = 1, 2, \dots, k$ by starting with $e_1 = f$ and taking the product $e_j = f \cup e_{j-1}$ for $j \geq 2$. We have that G is k -colorable if and only if $e_k(V) > 0$.

Surprise 3: Hamiltonian Path

Here we illustrate the third surprise, namely a randomized algorithm for the Hamiltonian path problem that runs in time $\mathcal{O}(1.66^n)$. This algorithm is due to Björklund.⁴ For ease of exposition, we restrict our consideration to bipartite graphs and obtain running time $\mathcal{O}(1.42^n)$. (The algorithm design here is also slightly different from Björklund’s original design; here we rely on reversal of a closed subwalk for cancellation of non-paths⁵ and, inspired by Cygan et al.,⁸ use the Isolation Lemma in place of polynomial identity testing.)

Let us return to the example in Figure 3. We observe that the graph is bipartite with $n = 8$, $V_1 = \{a, b, c, d\}$, and $V_2 = \{e, f, g, h\}$. As before, our task is to decide whether there exists a Hamiltonian path from vertex s to vertex t . Let us assume that $s = a$ and $t = h$.

Every walk of length $n - 1$ makes exactly n visits to vertices, where exactly $n/2$ visits are to vertices in V_1 because the graph is bipartite. Let us now *label* each of the $n/2$ visits to V_1 using an integer from $L = \{1, 2, \dots, n/2\}$. In particular, each walk has $(n/2)^{n/2}$ possible labelings, exactly $(n/2)!$ of which are *bijective*, that is, each label is used exactly once. For example, let us consider the labeled walk

$${}_1a {}_2e {}_3c {}_4b {}_5d {}_6f {}_7g {}_8h. \quad (7)$$

We observe that (7) is a bijectively labeled non-path.

Let us now partition the set of all labeled walks into two disjoint classes,

the “good” class and the “bad” class. A labeled walk is *good* if the labeling is bijective and the walk is a path. Otherwise a labeled walk is *bad*. We observe that the good class is non-empty if and only if the graph has a Hamiltonian path from s to t .

We now develop a randomized algorithm that decides whether the good class is nonempty. The key idea is to build a sieve for filtering labeled walks so that (a) the bad class is always filtered out and (b) a “witness” from the good class remains with fair probability whenever the good class is nonempty. Conceptually, it will be convenient to regard the sieve as a “bag” (multiset) to which we “hash” labeled walks so that upon termination each “bad” hash value will occur in the bag an even number of times, and each “good” hash value will occur exactly once.

Define the *hash* of a labeled walk to be the multiset that consists of all the elements visited by a walk, together with their labels (if any). For example, the hash value of (7) is

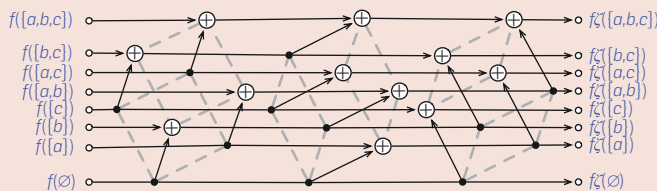
$$\{A, B, B, C, D, F, H, {}_1a, {}_2a, {}_3b, {}_4c, e, f, f, h\}. \quad (8)$$

In general, we cannot reconstruct a labeled walk from its hash value. However, every bijectively labeled path—that is, every good labeled walk—can be reconstructed from its hash value. Indeed, the vertices in a path are distinct, and the set of edges of a path determines the ordering of the vertices, which we know must start with s and end with t . Thus, each good labeled walk has a unique hash value.

Our next objective is to make sure that each hash value arising from a bad labeled walk gets inserted an even number of times into the sieve. Toward this end, there are two disjoint types of bad labeled walks, namely (a) bijectively labeled non-paths and (b) non-bijectively labeled walks.

Let us consider a bijectively labeled non-path W . We show that W can be paired with a bijectively labeled non-path W' with the same hash value. If we view W as a string, there is a minimal string prefix that contains a repeated vertex. Let us call the last vertex v in such a prefix the *first* repeated vertex in W . Let v be the first repeated vertex in W , and call the subwalk between the

Figure 7. Fast zeta transform for $n = 3$.



first two occurrences of v in W the *first closed subwalk* in W . For example, in (7) the first closed subwalk is $aAeC^1b^3DfBa^4$. There are two cases to consider in setting up the pairing, depending on whether the first repeated vertex in W is in V_1 or in V_2 .

If the first repeated vertex is in V_1 , let us define W' by transposing the labels of the first and last vertex in the first closed subwalk (that is, the first two occurrences of the first repeated vertex in W). For example, in the case of (7) we obtain

$$aAeC^1b^3DfBa^4BfFcHh. \quad (9)$$

Clearly, W and W' have the same hash value. Furthermore, because W is bijectively labeled, $W' \neq W$. Since $W'' = W$, we have a bijective pairing of bijectively labeled non-walks where the first repeated vertex is in V_1 .

If the first repeated vertex is in V_2 , let us *reverse* the first closed subwalk (also reversing the labels) in W to obtain the bijectively labeled non-path W' . For example,

$$aBfDbEgGcFfFcHh \quad (10)$$

gets paired with

$$aBfFcGgEbDfFcHh. \quad (11)$$

It is immediate that W and W' have the same hash value. We also observe that $W'' = W$ since two reversals restore the original bijectively labeled non-path. It remains to conclude that $W \neq W'$. Here it is not immediate that reversing the first closed subwalk will result in a different labeled walk. Indeed, the first closed subwalk may be a palindrome, such as eC^1b^2Ce in

$$aAeC^1b^2CeAaBfFcHh. \quad (12)$$

Fortunately, because of bijective labeling, the only possible pitfall is a palindrome of length 5 that starts at V_2 , visits a vertex in V_1 , and returns to the same vertex in V_2 . We can avoid such palindromes by keeping track of the last vertices visited by a partial walk, and hence assume that our labeled walks do not contain such palindromes, and consequently $W' \neq W$. Thus, the set of bijectively labeled non-paths partitions into disjoint

pairs $\{W, W'\}$, where each pair has the same hash value.

Next, let us consider a non-bijectively labeled walk W . Each such W *avoids at least one label* from the set of all labels L . In particular, if W avoids exactly a labels, there are exactly 2^a sets $A \subseteq L$ such that W avoids every label in A (and possibly some other labels outside A).

From the previous observations we now obtain the following high-level algorithm. For each subset $A \subseteq L$ in turn, we insert into the sieve the hash value of each labeled walk that avoids every label in A . After all subsets A have been considered, a hash value occurs with odd multiplicity in the sieve if and only if it originates from a good labeled walk.

A second key idea is now to implement the sieve at low level using what is essentially a layer of hashing so that the hash values—such as (8)—are not considered explicitly, but rather *by weight only*. That is, instead of sieving hash values explicitly, we sieve only their weights. In particular, at the start of the algorithm, let us associate an integer weight in the interval $1, 2, \dots, n(n+1)$ independently and uniformly at random to each of the $(n+1)n/2$ elements that may occur in a hash value. The *weight* of a hash value is the sum of the weights of its elements. When running the sieve, instead of tracking the (partial) walks and their (partial) hash values by dynamic programming, we only track the number of hash values of each weight. This enables us to process each fixed $A \subseteq L$ in time polynomial in n . The number of all sets $A \subseteq L$ is $2^{|L|} \leq 2^{n/2} < 1.42^n$. Thus, the total running time of the above procedure is $O(1.42^n)$. When the sieve terminates, we assert that the input graph has a Hamiltonian path if the counter for the number of hash values of at least one weight is odd; otherwise we assert that the graph has no Hamiltonian path.

To see that the presence of an odd counter implies the existence of a Hamiltonian path, observe that by our careful design, each bad hash value gets inserted into the sieve an even number of times, and in particular contributes an even increment to the counter corresponding to the

weight of the hash value. Thus, an odd counter can arise only if a good hash value was inserted into the sieve, that is, the graph has a Hamiltonian path.

Next, let us study the probability of a false negative, that is, all counters are even although the graph has a Hamiltonian path. Here it suffices to invoke the “Isolation Lemma” of Mulmuley et al.³⁰ which states that for any set family over a base set of m elements, if we assign a weight independently and uniformly at random from $1, 2, \dots, r$ to each element of the base set, there will be a unique set of the minimum weight in the family with probability at least $1 - m/r$. In particular, if we consider the set family of good hash values—indeed, each good hash value is a set—there is a unique such hash value of the minimum weight—and hence an odd counter in the sieve—with probability at least $1/2$.

We thus have a randomized algorithm for detecting Hamiltonian paths in bipartite graphs that runs in time $O(1.42^n)$, gives no false positives, and gives a false negative with probability at most $1/2$. (The algorithm could now be extended to graphs that are not bipartite with running time $O(1.66^n)$ by partitioning the vertices randomly into V_1 and V_2 and employing a bijective labeling also for the edges with both ends in V_2 .)

Conclusion

This article has highlighted three recent results in exact exponential algorithms, with the aim of illustrating the range of techniques that can be employed and the element of surprise in each case. In this regard, it is perhaps safe to say that the area is still in a state of flux, and with more research one can expect more positive surprises. Certainly, the authors do not mind to be labeled as optimists in this sense. We also hope the three highlighted results have illustrated perhaps the main reason why one wants to study algorithms that run in exponential time. That is, the study of exponential time algorithms is really a quest for understanding computation and the structure of computational problems, including pursuing the sometimes surprising connections uncovered in such a quest.

We conclude with three challenge problems, each of which at first sight appears quite similar to one of the three surprises we have covered in this article. Frustratingly enough, however, there has been no progress at all on these problems.

MAX-3-SAT. We have seen that MAX-2-SAT can be solved in time $\mathcal{O}(2^{m/3})$ essentially because of the existence of nontrivial algorithms for matrix multiplication. But no such tools are available when one considers instances with clauses of length 3 instead of length 2. The challenge is to find an algorithm that runs in time $\mathcal{O}((2 - \epsilon)^n)$ for MAX-3-SAT, where n is the number of variables and $\epsilon > 0$ is a constant independent of n .

Edge Coloring. The edge-coloring problem asks us to color the edges of a graph using the minimum number of colors such that the coloring is *proper*, that is, any two edges that share an endvertex must receive different colors. It is known that the number of colors required is either Δ or $\Delta + 1$, where Δ is the maximum degree of a vertex, and it is NP-complete to decide which of the two cases occurs.²⁰ For a graph G , the edge-coloring of G is equivalent to deciding whether the chromatic number of the line graph $L(G)$ of G is Δ or $\Delta + 1$, which implies that edge-coloring can be solved in time $2^m m^{\mathcal{O}(1)}$, where m is the number of edges in G . The challenge is to find an algorithm that runs in time $\mathcal{O}((2 - \epsilon)^m)$ where $\epsilon > 0$ is independent of m .

Traveling Salesman. While the Hamiltonian cycle problem can be solved in randomized time $\mathcal{O}(1.66^n)$, no such algorithm is known for the Traveling Salesman Problem with n cities and travel costs between cities that are nonnegative integers whose binary representation is bounded in length by a polynomial in n . The challenge is to find an algorithm that runs in time $\mathcal{O}((2 - \epsilon)^n)$ where $\epsilon > 0$ is independent of n .

Further Reading

Beyond the highlighted results in this article, the recent book of Fomin and Kratsch¹⁵ and the surveys of Woeginger^{38,39} provide a more in-depth introduction to exact exponential

algorithms. Dantsin and Hirsch⁹ survey algorithms for SAT, while Malik and Zhang²⁸ discuss the deployment of SAT solvers in practical applications. Husfeldt²¹ gives an introduction to applications of the principle of inclusion and exclusion in algorithmics. Flum and Grohe¹³ give an introduction to parameterized complexity theory and its connections to subexponential and exponential time complexity. Williams³⁷ relates improvements to exhaustive search with superpolynomial lower bounds in circuit complexity.

Acknowledgments

The authors would like to thank Andreas Björklund, Thore Husfeldt, Mikko Koivisto, and Dieter Kratsch for their comments that greatly helped to improve the exposition in this review. F.V.F. acknowledges the support of the European Research Council (ERC), grant Rigorous Theory of Preprocessing, reference 267959. P.K. acknowledges the support of the Academy of Finland, Grants 252083 and 256287. C

References

1. Bax, E.T. Inclusion and exclusion algorithm for the Hamiltonian path problem. *Inf. Process. Lett.* 47, 4 (1993), 203–207.
2. Bellman, R. *Dynamic Programming*, Princeton University Press, 1957.
3. Bellman, R. Dynamic programming treatment of the travelling salesman problem. *J. ACM* 9 (1962), 61–63.
4. Björklund, A. Determinant sums for undirected hamiltonicity. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS 2010)* (2010), IEEE, 173–182.
5. Björklund, A., Husfeldt, T., Kaski, P., Koivisto, M. Narrow sieves for parameterized paths and packings. *arXiv:1007.1161* (2010).
6. Björklund, A., Husfeldt, T., Koivisto, M. Set partitioning via inclusion–exclusion. *SIAM J. Comput.* 39, 2 (2009), 546–563.
7. Coppersmith, D., Winograd, S. Matrix multiplication via arithmetic progressions. *J. Symbolic Comput.* 9, 3 (1990), 251–280.
8. Cygan, M., Nederlof, J., Pilipczuk, M., Pilipczuk, M., van Rooij, J.M.M., Wojtaszczyk, J.O. Solving connectivity problems parameterized by treewidth in single exponential time. In *Proceedings of the 52nd Annual Symposium on Foundations of Computer Science* (2011), IEEE, 150–159.
9. Dantsin, E., Hirsch, E.A. Worst-case upper bounds. In *Handbook of Satisfiability*, volume 185 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2009, 403–424.
10. Downey, R.G., Fellows, M.R. *Parameterized Complexity*, Springer, 1999.
11. Eppstein, D. Small maximal independent sets and faster exact graph coloring. *J. Graph Algorithms Appl.* 7, 2 (2003), 131–140.
12. Eppstein, D. Quasiconvex analysis of multivariate recurrence equations for backtracking algorithms. *ACM Trans. Algorithms* 2, 4 (2006), 492–509.
13. Flum, J., Grohe, M. *Parameterized Complexity Theory*, Springer, 2006.
14. Fomin, F.V., Grandoni, F., Kratsch, D. A measure &

- conquer approach for the analysis of exact algorithms. *J. ACM* 56, 5 (2009).
15. Fomin, F.V., Kratsch, D. *Exact Exponential Algorithms*, Springer, 2010.
16. Fortnow, L. The status of the P versus NP problem. *Commun. ACM* 52, 9 (2009), 78–86.
17. Garey, M.R., Johnson, D.S. *Computers and Intractability, A Guide to the Theory of NP-Completeness*, W.H. Freeman and Company, 1979.
18. Golomb, S.W., Baumert, L.D. Backtrack programming. *J. ACM* 12 (1965), 516–524.
19. Held, M., Karp, R.M. A dynamic programming approach to sequencing problems. *J. Soc. Indust. Appl. Math.* 10 (1962), 196–210.
20. Holyer, I. The NP-completeness of edge-coloring. *SIAM J. Comput.* 10, 4 (1981), 718–720.
21. Husfeldt, T. Invitation to algorithmic uses of inclusion–exclusion. *arXiv:1105.2942* (2011).
22. Itai, A., Rodeh, M. Finding a minimum circuit in a graph. *SIAM J. Comput.* 7, 4 (1978), 413–423.
23. Karp, R.M. Dynamic programming meets the principle of inclusion and exclusion. *Oper. Res. Lett.* 1, 2 (1982), 49–51.
24. Knuth, D.E. *The Art of Computer Programming*, vol. 2: *Seminumerical Algorithms*, 3rd edn, Addison-Wesley, 1998.
25. Kohn, S., Gottlieb, A., Kohn, M. A generating function approach to the traveling salesman problem. In *Proceedings of the ACM Annual Conference (ACM 1977)* (1977), ACM Press, 294–300.
26. Koivisto, M. Optimal 2-constraint satisfaction via sum-product algorithms. *Inform. Process. Lett.* 98, 1 (2006), 24–28.
27. Lawler, E.L. A note on the complexity of the chromatic number problem. *Inf. Process. Lett.* 5, 3 (1976), 66–67.
28. Malik, S., Zhang, L. Boolean satisfiability: From theoretical hardness to practical success. *Commun. ACM* 52, 8 (2009), 76–82.
29. Moon, J.W., Moser, L. On cliques in graphs. *Israel J. Math.* 3 (1965), 23–28.
30. Mulmuley, K., Vazirani, U.V., Vazirani, V.V. Matching is as easy as matrix inversion. *Combinatorica* 7, 1 (1987), 105–113.
31. Niedermeier, R. *Invitation to Fixed-Parameter Algorithms*, Oxford University Press, 2006.
32. Strassen, V. Gaussian elimination is not optimal. *Numer. Math.* 13 (1969), 354–356.
33. Vassilevska Williams, V. Multiplying matrices faster than Coppersmith–Winograd. In *Proceedings of 44th ACM Symposium on Theory of Computing (STOC 2012)* (2012), ACM, 887–898.
34. Vazirani, V.V. *Approximation Algorithms*, Springer, 2001.
35. Walker, R.J. An enumerative technique for a class of combinatorial problems. In *Proceedings of Symposia in Applied Mathematics*, vol. 10, American Mathematical Society, 1960, 91–94.
36. Williams, R. A new algorithm for optimal 2-constraint satisfaction and its implications. *Theoret. Comput. Sci.* 348, 2–3 (2005), 357–365.
37. Williams, R. Improving exhaustive search implies superpolynomial lower bounds. In *Proceedings of 42nd ACM Symposium on Theory of Computing* (2010), ACM, 231–240.
38. Woeginger, G. Exact algorithms for NP-hard problems: a survey. In *Combinatorial Optimization – Eureka, You Shrink!* (2003), volume 2570 of *Lecture Notes in Computer Science*, Springer, 185–207.
39. Woeginger, G. Space and time complexity of exact algorithms: some open problems. In *Proceedings of the 1st International Workshop on Parameterized and Exact Computation* (2004), volume 3162 of *Lecture Notes in Computer Science*, Springer, 281–290.
40. Yates, F. *The Design and Analysis of Factorial Experiments*, Imperial Bureau of Soil Science, 1937.
41. Zykov, A.A. On some properties of linear complexes. *Mat. Sbornik N.S.* 24, 66 (1949), 163–188.

Fedor V. Fomin (fomin@ii.uib.no) is a professor in the Institutt for Informatikk, University of Bergen, Norway.

Petteri Kaski (petteri.kaski@aalto.fi) is an Academic Research Fellow in the Department of Information and Computer Science at Aalto University, Aalto, Finland.

research highlights

P. 90

Technical Perspective Video Quality Assessment in the Age of Internet Video

By David Oran

P. 91

Understanding the Impact of Video Quality on User Engagement

By Florin Dobrian, Asad Awan, Dilip Joseph, Aditya Ganjam,
Jibin Zhan, Vyas Sekar, Ion Stoica, and Hui Zhang

Technical Perspective

Video Quality Assessment in the Age of Internet Video

By David Oran

IN THE DAYS of one-size-fits-all video delivery, broadcasters and rebroadcasters (like satellite and cable companies) concerned themselves with maintaining very high delivery quality under conditions they mostly controlled, from the original released content to the users' eyeballs. Variations in delivered quality were, of course, present. Televisions differed in size and quality of electronics. Both analog and digital broadcast technologies had distance-based degradations and occasional partial or total failures. Nonetheless, users' expectations of these systems were set to "broadcast quality" and providers were held to that standard by both regulators and the content owners. Quality measurement for broadcast video, both via objective metrics and subjective assessment, are a mature field with generally well-understood methodology and agreed standards.

Once again, it appears that "the Internet changes everything." The variety of endpoints for consuming video has dramatically increased. The variety of formats and delivery protocols has exploded, with standards lagging invention by years. Consumer access has opened up beyond the broadcasters and licensees through the global availability of "over the top" video services like Netflix, Hulu, and YouTube. All of this is eerily reminiscent of the disruptive trajectory of Voice-over-IP (VoIP), which went from a hobbyist curiosity in the mid-1990s to a competitive mainstream commercial service in the early 2000s, to poised to completely replace the PSTN by the year 2020. The tipping point came around 2004, when the majority of global voice traffic was carried at least partially over the Internet.

As with VoIP, we are seeing consumers substantially changing their consumption behavior and their assessment of service quality. Convenience trumps quality. Quality expectations are lower, partially due to this, but also due to dramatically lower pricing for large

libraries of content. In this changed environment, what measures of quality are most relevant, and how are they obtained from consumers no longer tethered to a single delivery service? In the following paper, the authors study the relationship between various objective measures of video delivery quality and *user engagement*, which they propose as the overall measure of the effectiveness of a video service.

Prior to digital video services, which have the ability to easily instrument a large number of programmable endpoints, user engagement could be assessed only by expensive means and with small sample sets, as was done for decades by companies like Nielsen. Digital cable systems could get a rough measure of user engagement by logging channel-changing behaviors. More recently IPTV systems, which emulate classic over-the-air and cable broadcast systems using IP network technology, have collected both channel-change and video-quality information from endpoints. None of these systems, however, has amassed anywhere near the quantity and diversity of data available to the authors through the Conviva service, nor have they published extensive studies relating video quality measures to user engagement. For large-scale Internet video services, this study is groundbreaking and will provide a baseline set of measures for others to use in future work.

Some of the results are unsurprising. Stalls in video output (which the

authors call *buffering*) cause the most dramatic effects on user engagement, particularly for "live" content such as sports and news. At some level, video stalls are analogous to major failures or glitches of earlier systems, and cause the most profound dissatisfaction among users. Stalls are a widespread phenomenon new to Internet video, and stem from the time- and location-varying quality of Internet bandwidth, packet loss patterns, and congestive overload. The protocols used today for adaptive streaming of video are still primitive and the subject of both active research and standardization work. As they improve, the frequency and severity of stalls should decrease, no longer masking the magnitude of other effects such as encoding quality, variations in quality due to adaptation algorithms, and join time. Other results in this work must be seen in the light of the dominance of stalls compared to the other phenomena the authors evaluate.

Despite the ubiquity and popularity of Netflix and YouTube, we are still in the early days of the Internet video phenomenon. As with VoIP, users initially were delighted to the point of astonishment that the system worked at all, since it provided access to a significantly larger library of content, at lower cost, greater convenience, and on more devices than available in the "walled gardens" offered by broadcasters, satellite, and cable systems. Today we are in the hyper-growth phase, where users expect the systems to have high availability but not necessarily reach the quality of the legacy systems. It will not be long before quality as well is competitive with or even in some cases exceeds non-Internet systems.

We will need mature measurement methodologies to support these systems. This work is a good start. 

For large-scale Internet video services, this study is groundbreaking.

David Oran (oran@cisco.com) is a Cisco Fellow at Cisco Systems, Cambridge, MA.

© 2013 ACM 0001-0782/13/03

Understanding the Impact of Video Quality on User Engagement

By Florin Dobrian, Asad Awan, Dilip Joseph, Aditya Ganjam, Jibin Zhan, Vyas Sekar, Ion Stoica, and Hui Zhang

Abstract

As the distribution of video over the Internet is becoming mainstream, user expectation for high quality is constantly increasing. In this context, it is crucial for content providers to understand if and how video quality affects user engagement and how to best invest their resources to optimize video quality. This paper is a first step toward addressing these questions. We use a unique dataset that spans different content types, including short video on demand (VoD), long VoD, and live content from popular video content providers. Using client-side instrumentation, we measure quality metrics such as the join time, buffering ratio, average bitrate, rendering quality, and rate of buffering events. We find that the percentage of time spent in buffering (buffering ratio) has the largest impact on the user engagement across all types of content. However, the magnitude of this impact depends on the content type, with live content being the most impacted. For example, a 1% increase in buffering ratio can reduce user engagement by more than 3 min for a 90-min live video event.

1. INTRODUCTION

Video content already constitutes a dominant fraction of Internet traffic today and several analysts forecast that this contribution is set to increase in the next few years.^{1, 15} This trend is fueled by the ever decreasing cost of content delivery and the emergence of new subscription- and ad-based business models. Premier examples are Netflix which now has reached 20 million US subscribers, and Hulu which distributes over one billion videos per month.

As Internet video goes mainstream, users' expectations for quality have dramatically increased; for example, when viewing content on TV screens anything less than "SD" quality (i.e., 480p) is not acceptable. In the spirit of Herbert Simon's articulation of attention economics, the overabundance of video content increases the onus on content providers to maximize their ability to attract users' attention.¹⁸ Thus, it becomes critical to systematically understand the interplay between video quality and user engagement. This knowledge can help providers to better invest their network and server resources toward optimizing the quality metrics that really matter.² However, our understanding of many key questions regarding the impact of video quality on user engagement "in the wild" is limited on several fronts:

1. Does poor video quality reduce user engagement? And by how much?
2. Do different quality metrics vary in the degree in which they impact the user engagement?
3. Does the impact of the quality metrics differ across content genres and across different granularities of user engagement?

This paper is a first step toward answering these questions. We do so using a unique dataset of *client-side* measurements obtained over 2 million unique video viewing sessions from over 1 million viewers across popular content providers. Using this dataset, we analyze the impact of video quality on user engagement along three dimensions:

- **Different quality metrics:** We capture characteristics of the start up latency, the rate at which the video was encoded, how much and how frequently the user experienced a buffering event, and what was the observed quality of the video rendered to the user.
- **Multiple timescales of engagement:** We quantify the user engagement at two levels: per-view (i.e., a single video being watched) and per-viewer (i.e., aggregated over all views for a specific user).
- **Different types of content:** We partition our data based on video type and length into short VoD, long VoD, and live, to represent the three broad types of video content being served today.

To identify the critical quality metrics and to understand the dependencies among these metrics, we employ well-known techniques such as correlation and information gain. We also augment this qualitative analysis with regression techniques to quantify the impact. Our main observations are

- The percentage of time spent in buffering (buffering ratio) has the largest impact on the user engagement across all types of content. However, this impact is quantitatively different for different content types, with live content being the most impacted. For a highly popular 90-min soccer game, for example, an increase in the buffering ratio of only 1% can lead to more than 3 min of reduction in the user engagement.

The original version of this paper with the same title was published in *ACM SIGCOMM*, 2011.

- The average bitrate at which the content is streamed has a significantly higher impact for live content than for VoD content.
- The quality metrics affect not only the per-view engagement but also the number of views watched by a viewer over a time period. Further, the join time has greater impact on viewer-level engagement.

These results have important implications on how content providers can best use their resources to maximize user engagement. Reducing the buffering ratio can increase the engagement for all content types, minimizing the rate of buffering events can improve the engagement for long VoD and live content, and increasing the average bitrate can increase the engagement for live content. However, there are also trade-offs between the buffering and the bitrate that we should take into account. Our ultimate goal is to use such measurement-driven insights so that content providers, delivery systems, and end users can objectively evaluate and improve Internet video delivery. The insights we present are a small, but significant, first step toward realizing this vision.

2. PRELIMINARIES AND DATASETS

We begin this section with an overview of how our dataset was collected. Then, we scope the three dimensions of the problem space: user engagement, video quality metrics, and types of video content.

2.1. Data collection

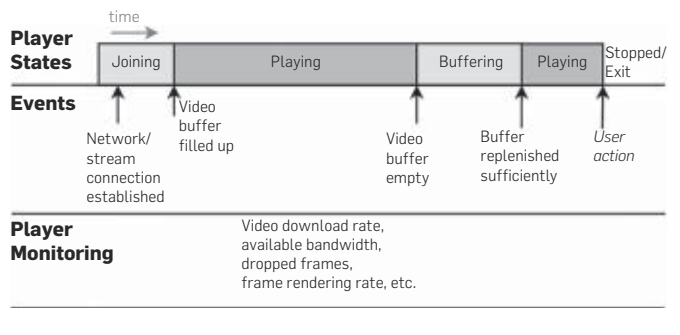
We have implemented a highly scalable and available real-time data collection and processing system. The system consists of two parts: (a) a client-resident instrumentation library in the video player and (b) a data aggregation and processing service that runs in data centers. Our client library gets loaded when Internet users watch video on our affiliates' sites and monitors fine-grained events and player statistics. This library collects high fidelity raw data to generate higher level information on the client side and transmits these in real time with minimal overhead. We collect and process 0.5TB of data on average per day from various affiliates over a diverse spectrum of end users, video content, Internet service providers, and content delivery networks.

Video player instrumentation: Figure 1 illustrates the lifetime of a video session as observed at the client. The video player goes through multiple states (connecting and joining, playing, paused, buffering, stopped). For example, the player goes to paused state if the user presses the pause button on the screen, or if the video buffer becomes empty then the player goes into buffering state. By instrumenting the client, we can observe all player states and events and also collect statistics about the playback quality.

2.2. Engagement and quality metrics

Qualitatively, engagement is a reflection of user involvement and interaction. While there are many ways in which we can define engagement (e.g., user-perceived satisfaction with the content or willingness to click advertisements), in this study we focus on objectively measurable metrics of engagement at two levels:

Figure 1. An illustration of a video session lifetime and associated video player events.



1. **View level:** A user watching a single video continuously is a view. For example, this could be watching a movie trailer clip, an episode of a TV serial, or a football game. The view-level engagement metric of interest is the *play time*—the duration of the viewing session. Note that we do not count ads in the stream as separate views; they are part of the actual content.
2. **Viewer level:** To capture the aggregate experience of a single viewer (an end user identified by a unique system-generated client ID), we study the viewer-level engagement metrics for each unique viewer. The two metrics we use are the number of views and the total play time across all videos watched by the viewer.

For completeness, we briefly describe the five industry-standard video quality metrics we use in this study²:

1. **Join time (*JoinTime*):** This represents the duration from the time at which the player initiates a connection to a video server till the time sufficient player video buffer has filled up and the player starts rendering video frames and moves to playing state. In Figure 1, this is the length of the joining state.
2. **Buffering ratio (*BufRatio*):** This is the fraction of the total session time (i.e., playing plus buffering time) spent in buffering. This is an aggregate metric that can capture periods of long video “freeze” observed by the user. As illustrated in Figure 1, the player goes into a buffering state when the video buffer becomes empty and moves out of buffering (back to playing state) when the buffer is replenished.
3. **Rate of buffering events (*RateBuf*):** *BufRatio* does not capture the frequency of induced interruptions. For example, a video session that experiences “video stuttering,” where each interruption is small but the total number of interruptions is high, might not have a high buffering ratio, but may be just as annoying to a user. Thus, we use the rate of buffering events $\frac{\# \text{ buffering events }}{\text{session duration}}$.
4. **Average bitrate (*AvgBitrate*):** A single video session can have multiple bitrates if the video player can switch between different bitrate streams. Such bitrate adaptation logic is widely deployed in commercial players today. This metric is simply the

average of the bitrates played weighted by the duration each bitrate is played.

5. **Rendering quality (*RendQual*)**: Rendering rate (frames per second) is central to user's visual perception. This rate may drop because of either CPU effects (e.g., player may drop frames if the CPU is overloaded) or network effects (e.g., congestion causes the buffer to become empty). To normalize the metric across videos that have different encoded frame rates, we define rendering quality as the ratio of the rendered frames per second to the encoded frames per second of the stream.

2.3. Dataset

We collect close to 4TB of data each week. On average, 1 week of our data captures measurements of over **300 million views** watched by about **100 million unique viewers** across all of our affiliate content providers. The analysis in this paper is based on the data collected from five of our affiliates in the fall of 2010. These providers appear in the Top-500 most popular sites and serve a large volume of video content, thus providing a representative view of Internet video quality and engagement.

We organize the data into three content types and within each content type we choose two datasets from different providers. We choose diverse providers in order to rule out biases induced by the particular provider or the player-specific optimizations and algorithms they use. For live content, we use additional data from the largest live streaming sports event of 2010: the FIFA World Cup. Table 1 summarizes the number of unique videos and viewers for each dataset, described below. To ensure that our analysis is statistically meaningful, we only select videos that have at least 1000 views over the week-long period.

- **Long VoD**: Long VoD clips are videos that are at least 35 min and at most 60 min in duration. They are typically full episodes of TV shows. The two long VoD datasets are labeled as *LvodA* and *LvodB*.
- **Short VoD**: We categorize video clips as short VoD if the video length is 2–5 min. These are often trailers, short interviews, short skits, and news clips. The two short VoD datasets are labeled as *SvodA* and *SvodB*.
- **Live**: Sports events and news feeds are typically delivered as live video streams. There are two key differences between the VoD-type content and live streams. First, the

client buffers in this case are sized such that the viewer does not lag more than a few seconds behind the video source. Second, all viewers are roughly synchronized in time. The two live datasets are labeled *LiveA* and *LiveB*. As a special case study, the dataset *LiveWC* corresponds to the three of the final World Cup games with almost a million viewers per game on average.

3. ANALYSIS TECHNIQUES

In this section, we show preliminary measurements to motivate the types of questions that we want to answer and briefly describe the analysis techniques we use.

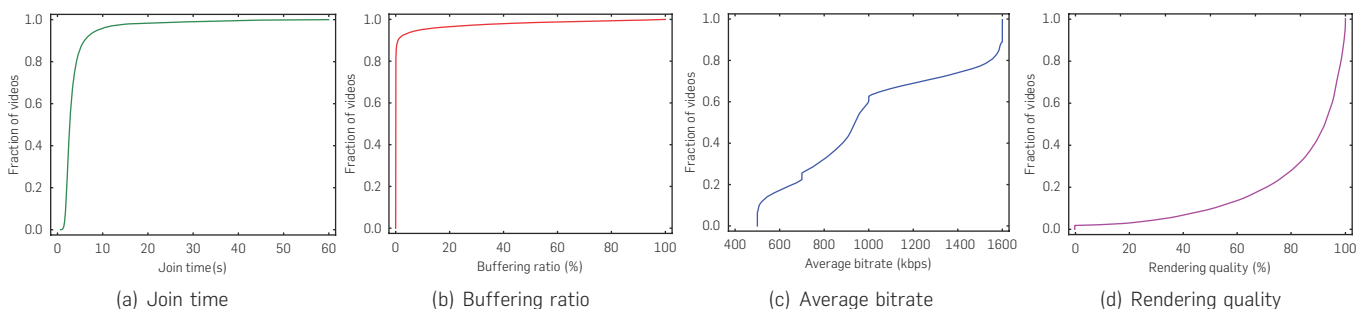
Overview: Figure 2 shows the cumulative distribution functions (CDF) of four quality metrics for dataset *LvodA*. We see that most viewing sessions experience very good quality, that is, have low *BufRatio*, low *JoinTime*, and relatively high *RendQual*. At the same time, however, the number of views that suffer from quality issues is not trivial—7% experience *BufRatio* larger than 10%, 5% have *JoinTime* larger than 10s, and 37% have *RendQual* lower than 90%. Finally, only a small fraction of views receive the highest bitrate. Given that a significant number of views experience quality issues, content providers would naturally like to know if (and by how much) improving their quality could have potentially increased the user engagement.

As an example, we consider one video object each from *LiveA* and *LvodA*, bin the different sessions based on the quality metrics, and calculate the average play time for each bin in Figures 3 and 4. These figures visually confirm that *quality matters*. At the same time, these initial visualizations also give rise to several questions:

Table 1. Summary of the datasets in our study.

Dataset	Number of videos	Number of viewers (100K)
<i>LiveA</i>	107	4.5
<i>LiveB</i>	194	0.8
<i>LvodA</i>	115	8.2
<i>LvodB</i>	87	4.9
<i>SvodA</i>	43	4.3
<i>SvodB</i>	53	1.9
<i>LiveWC</i>	3	29

Figure 2. Cumulative distribution functions for four quality metrics for dataset *LvodA*.



- How do we identify which metrics matter the most?
- Are these quality metrics independent or is the observed relationship between the engagement and the quality metric M due to a hidden relationship between M and a more critical metric M' ?
- How do we quantify how important a quality metric is?
- What causes the seemingly counterintuitive behaviors? For example, *RendQual* is negatively correlated (Figure 4(d)), while the *AvgBitrate* shows non-monotone relationships (Figure 3(c)).

To address the first two questions, we use the well-known concepts of correlation and information gain. To measure the quantitative impact, we also use linear-regression-based models for the most important metric(s). Finally, we use domain-specific insights and controlled experiments to explain the anomalous observations. Next, we briefly describe the statistical techniques we employ.

Correlation: To avoid making assumptions about the nature of the relationships between the variables, we choose the Kendall correlation instead of the Pearson correlation. The Kendall correlation is a rank correlation that does not make any assumption about the underlying distributions, noise, or the nature of the relationships. (Pearson correlation assumes that the noise in the data is Gaussian and that the relationship is roughly linear.)

Given the raw data—a vector of (x, y) values where each x is the measured quality metric and y the engagement metric (play time or number of views)—we bin it based on the value of the quality metric. We choose bin sizes that are appropriate for each quality metric of interest: for *JoinTime*, we use

0.5s intervals, for *BufRatio* and *RendQual* we use 1% bins, for *RateBuf* we use 0.01/min sized bins, and for *AvgBitrate* we use 20 kbps-sized bins. For each bin, we compute the empirical mean of the engagement metric across the sessions/viewers that fall in the bin.

We compute the Kendall correlation between the mean-per-bin vector and the values of the bin indices. We use this binned correlation metric for two reasons. First, we observed that the correlation coefficient was biased by a large mass of users that had high quality but very low play time, possibly because of low user interest. Our goal in this paper is not to study user interest. Rather, we want to understand how the quality impacts user engagement. To this end, we look at the average value for each bin and compute the correlation on the binned data. The second reason is scale. Computing the rank correlation is expensive at the scale of analysis we target; binned correlation retains the qualitative properties at much lower computation cost.

Information Gain: Correlation is useful when the relationship between the variables is roughly monotone increasing or decreasing. As Figure 3(c) shows, this may not hold. Furthermore, we want to move beyond analyzing a single quality metric. First, we want to understand if a pair (or a set) of quality metrics are complementary or if they capture the same effects. As an example, consider *RendQual* in Figure 3; *RendQual* could reflect either a network issue or a client-side CPU issue. Because *BufRatio* is also correlated with *PlayTime*, we may suspect that *RendQual* is mirroring the same effect. Identifying and uncovering these hidden relationships, however, is tedious. Second, content providers may want to know the top- k metrics that they should

Figure 3. Qualitative relationships between four quality metrics and the play time for one video in Lvoda.

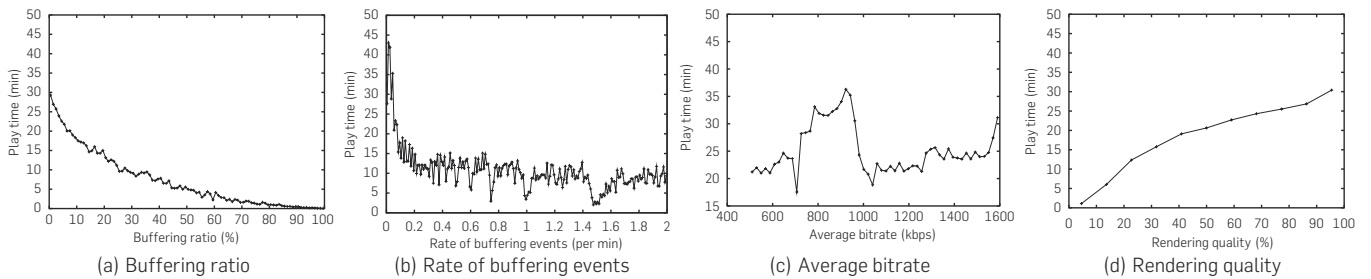
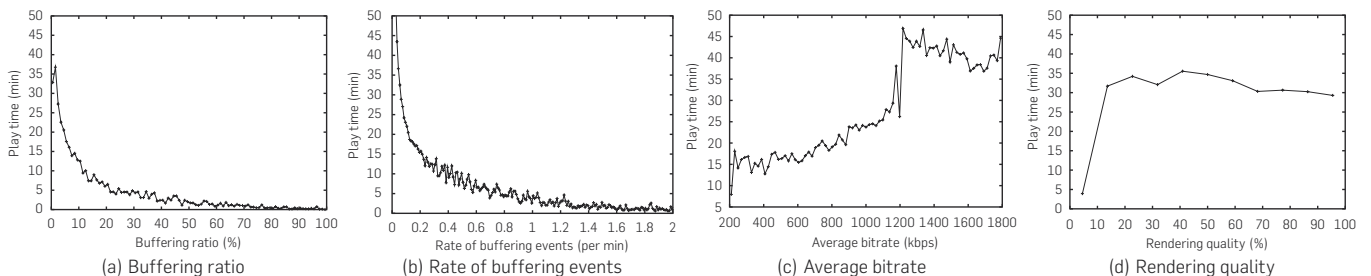


Figure 4. Qualitative relationships between four quality metrics and the play time for one video in LiveA.



optimize to improve user engagement.

To this end, we augment the correlation analysis using *information gain*,¹⁶ which is based on the concept of entropy. Intuitively, this metric quantifies how our knowledge of a variable X reduces the uncertainty in another variable Y ; for example, what does knowing the *AvgBitrate* or *BufRatio* “inform” us about the *PlayTime* distribution? We use a similar strategy to bin the data and for the *PlayTime*, we choose different bin sizes depending on the duration of the content.

Note that these analysis techniques are complementary. Correlation provides a first-order summary of monotone relationships between engagement and quality. The information gain can corroborate the correlation or augment it when the relationship is not monotone. Further, it extends our understanding to analyze interactions across quality metrics.

Regression: Kendall correlation and information gain are largely qualitative measures. It is also useful to understand the quantitative impact; for example, what is the expected increase in engagement if we improve a specific quality metric by a given amount? Here, we rely on regression. However, as the visualizations show, the relationships between the quality metrics and the engagement are not obvious and many metrics have intrinsic dependencies. Thus, directly applying regression techniques may not be meaningful. As a simpler and more intuitive alternative, we use linear regression to quantify the impact of specific ranges of the most critical quality metric. However, we do so only after visually confirming that the relationship is roughly linear over this range so that the linear data fit is easy to interpret.

4. ENGAGEMENT ANALYSIS

We begin by analyzing engagement at the per-view level, where our metric of interest is *PlayTime*. We begin with long VoD content, then proceed to live and short VoD content. In each case, we compute the binned correlation and information gain per video and then look at the distribution of the coefficients across all videos. Having identified the most critical metric(s), we quantify the impact of improving this quality using a linear regression model over a specific range of the quality metric.

At the same time, content providers also want to understand if good quality improves customer retention or if it encourages users to try more videos. Thus, we also analyze the user engagement at the viewer level by considering the number of views per viewer and the total play time across all videos watched by the viewer in a 1-week interval.

4.1. Long VoD content

Figure 5 shows the absolute and signed values of the correlation coefficients for *LvodA* to show the magnitude and the nature (increasing or decreasing) of the correlation. We summarize the median values for both datasets in Table 2 and find that the results are consistent for the common quality metrics *BufRatio*, *JoinTime*, and *RendQual*, confirming that our observations are not unique to a specific provider.

The result shows that *BufRatio* has the strongest correlation with *PlayTime*. Intuitively, we expect a higher *BufRatio*

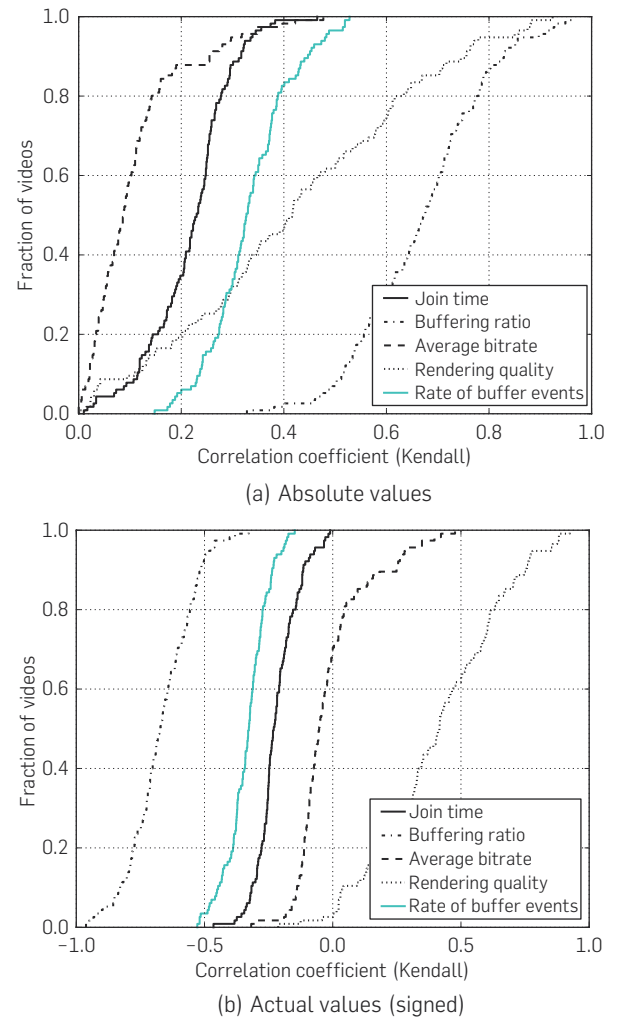
to decrease *PlayTime* (i.e., more negative correlation) and a higher *RendQual* to increase *PlayTime* (i.e., a positive correlation). Figure 5(b) confirms this intuition regarding the nature of these relationships. We also notice that *JoinTime* has little impact on the play duration.

Next, we use the univariate information gain analysis to corroborate and complement the correlation results. In Figure 6, the relative order between *RateBuf* and *BufRatio* is

Table 2. Median values of the Kendall rank correlation coefficients for *LvodA* and *LvodB*. We do not show *AvgBitrate* and *RateBuf* for *LvodB* because the player did not switch bitrates or gather buffering event data. For the remaining metrics, the results are consistent with dataset *LvodA*.

Quality Metric	Correlation coefficient	
	<i>LvodB</i>	<i>LvodA</i>
<i>JoinTime</i>	-0.17	-0.23
<i>BufRatio</i>	-0.61	-0.67
<i>RendQual</i>	0.38	0.41

Figure 5. Distribution of the Kendall rank correlation coefficients between quality metrics and play time for *LvodA*.



reversed compared to Figure 5. The reason is that most of the probability mass is in the first bin (0–1% *BufRatio*) and the entropy here is the same as the overall distribution (not shown). Consequently, the information gain for *BufRatio* is low; *RateBuf* does not suffer this problem and has higher information gain. Curiously, we see that *AvgBitrate* has high information gain even though its correlation with *PlayTime* is very low; we revisit this later in the section.

So far we have looked at each quality metric in isolation. A natural question then is whether two or more metrics when combined together yield new insights that a single metric does not provide. However, this may not be the case if the metrics are themselves interdependent. For example, *BufRatio* and *RendQual* may be correlated with each other; thus knowing that both are correlated with *PlayTime* does not add new information. Thus, we consider the distribution of the bivariate relative information gain values in Figure 7. For clarity, rather than showing all combinations, for each metric we include the bivariate combination with the highest relative information gain. We see that the

Figure 6. Distribution of the univariate gain between the quality metrics and play time for *Lvoda*.

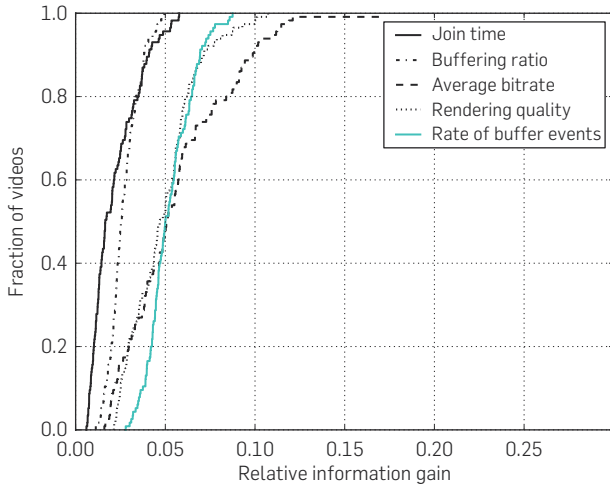
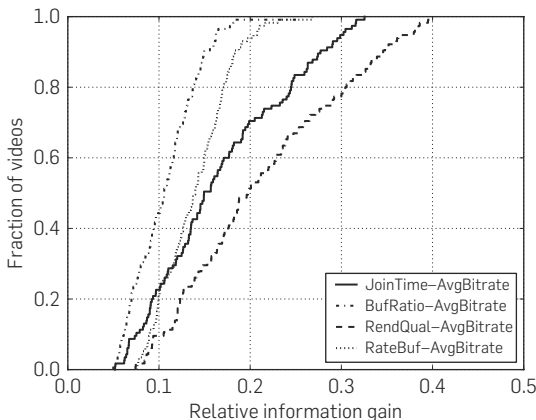


Figure 7. Distribution of the bivariate (relative) information gain for *Lvoda*. For brevity, we only show the best bivariate combinations.



combination with the *AvgBitrate* provides the highest bivariate information gain. Even though *BufRatio*, *RateBuf*, and *RendQual* had strong correlations in Figure 5, combining them does increase the information gain suggesting that they are interdependent.

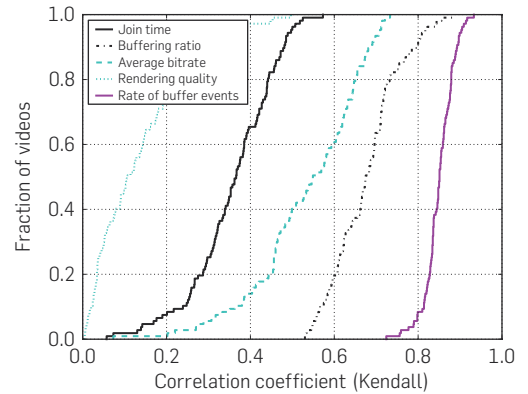
Surprising behavior in *AvgBitrate*: We noticed that *AvgBitrate* has low correlation but high information gain in the univariate and bivariate analysis. This is related to our earlier observation in Figure 3. The relationship between *PlayTime* and *AvgBitrate* is not monotone; it peaks between 800 and 1000 Kbps, low on either side of this region, and increases slightly at the highest rate. Because of this non-monotone relationship, the correlation is low.

However, knowing the value of *AvgBitrate* allows us to predict the *PlayTime* and thus there is a non-trivial information gain. This still leaves open the issue of low *PlayTime* in the 1000–1600 kbps band. This range corresponds to clients that observe many bitrate switches because of buffering induced by poor network conditions. Thus, the *PlayTime* is low here as a result of buffering, which we already observed to be the most critical factor.

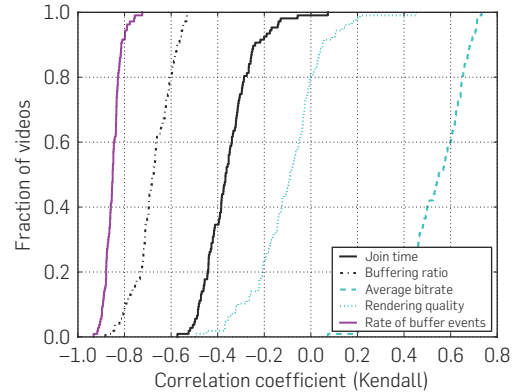
4.2. Live content

Figure 8 shows the distribution of the correlation coefficients for dataset *LiveA*, and we summarize the median

Figure 8. Distribution of the Kendall rank correlation coefficient between the quality metrics and play time, for dataset *LiveA*.



(a) Absolute values



(b) Actual values (signed)

values for the two datasets in Table 3. We notice one key difference with respect to the *Lvoda* results: *AvgBitrate* is more strongly correlated for live content. Similar to dataset *Lvoda*, *BufRatio* is strongly correlated, while *JoinTime* is weakly correlated.

For both long VoD and live content, *BufRatio* is a critical metric. Interestingly, for live, we see that *RateBuf* has a much stronger negative correlation with *PlayTime*. This suggests that the Live users are more sensitive to each buffering event compared to the Long VoD audience. Investigating this further, we find that the average buffering duration is much smaller for long VoD (3 s), compared to live (7 s). That is, each buffering event in the case of live content is more disruptive. Because the buffer sizes in long VoD are larger, the system fares better in face of fluctuations in link bandwidth. Furthermore, the system can be more proactive in predicting buffering and hence preventing it by switching to another server, or switching bitrates. Consequently, there are fewer and shorter buffering events for long VoD.

Information gain analysis reconfirms that *AvgBitrate* is a critical metric and that *JoinTime* is less critical for Live content (not shown). The bivariate results (not shown for brevity) mimic the same effects as those depicted in Figure 7, where the combination with *AvgBitrate* has the largest information gains.

Surprising behavior with *RendQual*: Figure 4(d) shows the counter-intuitive effect where *RendQual* was negatively correlated with *PlayTime* for live content. The above results for the *LiveA* and *LiveB* datasets confirm that this is not an anomaly specific to one video but a more pervasive phenomenon. Investigating this further, we found a surprisingly large fraction of viewers with low rendering quality and high play time. Furthermore, the *BufRatio* values for these users were also very low. In other words, these users see a drop in *RendQual* even without any network issues but continue to view the video.

We hypothesized that this effect arises out of a combination of user behavior and player optimizations. Unlike long VoD viewers, live video viewers may run the video player in background or minimize the browser (and maybe listening to the commentary). In this case, the player may try to reduce the CPU consumption by decreasing the frame rendering rate. To confirm this hypothesis, we replicated this behavior in a controlled setup and found that the player drops the *RendQual* to 20%. Interestingly, the *PlayTime* peak in Figure 4(d) also occurs at 20%. These suggest that the

anomalous relationship is due to player optimizations when users play the video in the background.

Case study with high impact events: One concern for content providers is whether the observations from typical videos can be applied to “high impact” events (e.g., Olympics¹⁰). To address this concern, we consider the *LiveWG* dataset. We focus here on *BufRatio* and *AvgBitrate*, which we observed as the most critical metrics for live content in the previous discussion. Figure 9 shows that the results for *LiveWC1* roughly match the results for *LiveA* and *LiveB*. We also confirmed that the coefficients for *LiveWG2* and *LiveWC3* are identical. These results suggest that our observations apply to such events as well.

4.3. Short VoD content

Finally, we consider the short VoD category. For both datasets *Svoda* and *Svodb*, the player uses a discrete set of 2–3 bitrates without switching and was not instrumented to gather buffering event data. Thus, we do not show the *AvgBitrate* (correlation is not meaningful on two points) and *RateBuf*. Table 4 summarizes the median values for both datasets. We notice similarities between long and short VoD: *BufRatio* and *RendQual* are the most critical metrics. As before, *JoinTime* is weakly correlated. The information gain results for short VoD largely mirror the results from the correlation analysis and we do not show these.

4.4. Quantitative analysis

As our measurements show, the interaction between the *PlayTime* and the quality metrics can be quite complex.

Figure 9. Impact of two quality metrics for *LiveWC1*, one of the three final games from the 2010 FIFA World Cup. A linear data fit is shown over the 0–10% subrange of *BufRatio*.

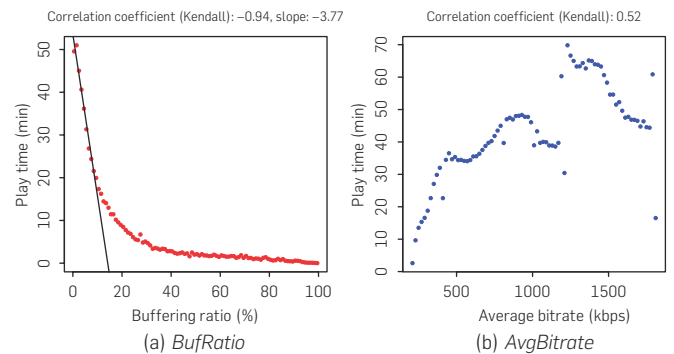


Table 3. Median values of the Kendall rank correlation coefficients for *LiveA* and *LiveB*. We do not show *AvgBitrate* and *RateBuf* because they do not apply for *LiveB*. For the remaining metrics the results are consistent with dataset *LiveA*.

Quality Metric	Correlation coefficient	
	<i>LiveB</i>	<i>LiveA</i>
<i>JoinTime</i>	-0.49	-0.36
<i>BufRatio</i>	-0.81	-0.67
<i>RendQual</i>	-0.16	-0.09

Table 4. Median values of the Kendall rank correlation coefficients for *Svoda* and *Svodb*. We do not show *AvgBitrate* and *RateBuf* because they do not apply here.

Quality Metric	Correlation coefficient	
	<i>Svodb</i>	<i>Svoda</i>
<i>JoinTime</i>	0.06	0.12
<i>BufRatio</i>	-0.53	-0.38
<i>RendQual</i>	0.34	0.33

Thus, we avoid black-box regression models and restrict our analysis to the most critical metric (*BufRatio*) and only apply regression to the 0–10% range of *BufRatio* after visually confirming that this is roughly a linear relationship.

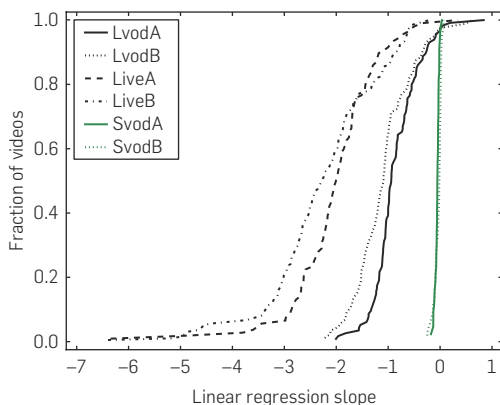
We notice that the distribution of the linear-fit slopes is very similar within the same content type in Figure 10. The median magnitudes of the slopes are one for long VoD, two for live, and close to zero for short VoD. That is, *BufRatio* has the strongest quantitative impact on live, then on long VoD, and then on short VoD. Figure 9 also includes linear data fits on the 0–10% subrange for *BufRatio* for the *LiveWG* data. These show that, within the selected subrange, a 1% increase in *BufRatio* can reduce the average play time by more than 3 min (assuming a game duration of 90 min). In other words, providers can increase the average user engagement by more than 3 min by investing resources to reduce *BufRatio* by 1%. Note that the 3 min drop is not from the 90-min content time but from expected view time which is around 40 min; that is, engagement drops by roughly 7.5% ($\frac{3}{40}$).

4.5. Viewer-level engagement

At the viewer level, we look at the aggregate *number of views* and *play time* per viewer across all objects irrespective of that video's popularity. For each viewer, we correlate the average value of each quality metric across different views with these two aggregate engagement metrics.

Figure 11 visually confirms that the quality metrics also impact the number of views. One interesting observation with *JoinTime* is that the number of views increases in the range 1–15s before starting to decrease. We also see a similar effect for *BufRatio*, where the first few bins have fewer total views. This effect does not, however, occur for the total play time. We speculate that this is an effect of user interest. Many users have very good quality but little interest in the content; they sample the content and leave without returning. Users who are actually interested in the content are more tolerant of longer join times (and buffering). However, the tolerance drops beyond a certain point (around 15s for *JoinTime*).

Figure 10. CDF of the linear-fit slopes between *PlayTime* and the 0–10% subrange of *BufRatio*.



The values of the correlation coefficients are qualitatively consistent across the different datasets (not shown) and also similar to the trends we observed at the view level. The key difference is that while *JoinTime* has relatively little impact at the view level, it has a more pronounced impact at the viewer level. This has interesting system design implications. For example, a provider may decide to increase the buffer size to alleviate buffering issues. However, increasing buffer size can increase *JoinTime*. The above result shows that doing so without evaluating the impact at the viewer level may reduce the likelihood of a viewer visiting the site again.

5. RELATED WORK

Content popularity: There is an extensive literature on modeling content popularity and its implications for caching (e.g., Cheng et al.,⁶ Yu et al.,¹² and Huang et al.¹⁴). While our analysis of the impact on quality on engagement is orthogonal, one interesting question is if the impact of quality differs across popularity segments, for example, is niche content more likely to be affected by poor quality?

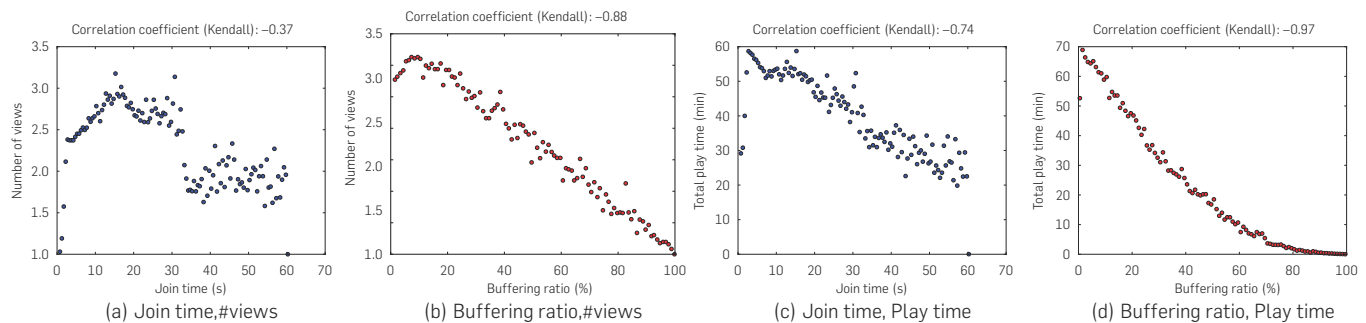
User behavior: Yu et al. observe that many users have small session times as they “sample” a video and leave.¹² Removing this potential bias was one of the motivations for our binned correlation analysis. Other researchers have studied channel switching in IPTV (e.g., Cha et al.⁸) and seek-pause-forward behaviors in streaming systems (e.g., Costa et al.⁷). These highlight the need to understand user behavior to provide better context for the measurements similar to our browser minimization scenario for live content.

Measurements of video delivery systems: The research community has benefited immensely from measurement studies of deployed VoD and streaming systems using both “black-box” inference (e.g., Gill et al.,⁴ Hei et al.,¹³ and Saroiu et al.¹⁷) and “white-box” measurements (e.g., Chang et al.,⁹ Yin et al.,¹⁰ and Sripanidkulchai et al.¹⁹). Our work follows in this rich tradition of measuring real deployments. At the same time, we have taken a significant first step to systematically analyze the impact of the video quality on user engagement.

User perceived quality: Prior work has relied on controlled user studies to capture user perceived quality indices (e.g., Gulliver and Ghinea¹¹). The difference in our work is simply an issue of timing and scale. Internet video has only recently attained widespread adoption; revisiting user engagement is ever more relevant now than before. Also, we rely on real-world measurements with millions of viewers rather than small-scale controlled experiments with a few users.

Engagement in other media: Analysis of understanding user engagement appears in other content delivery mechanisms as well: impact of Website loading times on user satisfaction (e.g., Bouch et al.⁵), impact of quality metrics such as bitrate, jitter, and delay on call duration in VoIP (e.g., Chen et al.³), among others. Our work is a step toward obtaining similar insights for Internet video delivery.

Figure 11. Visualizing the impact of *JoinTime* and *BufRatio* on the number of views and play time for *Lvoda*.



6. REFLECTIONS

The findings presented in this paper are the result of an iterative process that included more false starts and misinterpretations than we care to admit. We conclude with two cautionary lessons we learned that apply more broadly to future studies of this scale.

The need for complementary analysis: For the long VoD case, we observed that the correlation coefficient for the average bitrate was weak, but the univariate information gain was high. The process of trying to explain this discrepancy led us to visualize the behaviors. In this case, the correlation was weak because the relationship was not monotone. The information gain, however, was high because the intermediate bins near the natural modes had significantly lower engagement and consequently low entropy in the play time distribution. This observation guided us to a different phenomenon, sessions that were forced to switch rates because of poor network quality. If we had restricted ourselves to a purely correlation-based analysis, we might have missed this effect and incorrectly inferred that *AvgBitrate* was not important. This highlights the value of using multiple views from complementary analysis techniques in dealing with large datasets.

The importance of context: Our second lesson is that while statistical techniques are excellent tools, they need to be used with caution and we need to take the results of these analyses together with the context of the user and system-level factors. For example, naively acting on the observation that the *RendQual* quality is negatively correlated for live content can lead to an incorrect understanding of its impact on engagement. As we saw, this is an outcome of user behavior and player optimizations. This highlights the importance of backing the statistical analysis with domain-specific insights and controlled experiments to replicate the observations. C

References

1. Cisco forecast, http://blogs.cisco.com/sp/comments/cisco_visual_networking_index_forecast_annual_update/.
2. Driving engagement for online video, <http://registration.digitallyspeaking.com/akamai/mddec10/registration.html?b=videonuze>.
3. Chen, K., Huang, C., Huang, P., Lei, C. Quantifying Skype user satisfaction. In *Proceedings of SIGCOMM* (2006).
4. Gill, P., Arlitt, M., Li, Z., Mahanti, A. YouTube traffic characterization: A view from the edge. In *Proceedings of IMC* (2007).

5. Bouch, A., Kuchinsky, A., Bhatti, N. Quality is in the eye of the beholder: Meeting users' requirements for Internet quality of service. In *Proceedings of CHI* (2000).
6. Cheng, B., Liu, X., Zhang, Z., Jin, H. A measurement study of a peer-to-peer video-on-demand system. In *Proceedings of IPTPS* (2007).
7. Costa, C., Cunha, I., Borges, A., Ramos, C., Rocha, M., Almeida, J., Ribeiro-Neto, B. Analyzing client interactivity in streaming media. In *Proceedings of WWW* (2004).
8. Cha, M., Rodriguez, P., Crowcroft, J., Moon, S., Amatriain, X. Watching television over an IP network. In *Proceedings of IMC* (2008).
9. Chang, H., Jamin, S., Wang, W. Live streaming performance of the Zattoo network. In *Proceedings of IMC* (2009).
10. Yin, H., Liu, X., Qiu, F., Xia, N., Lin, C., Zhang, H., Sekar, V., Min, G. Inside the bird's nest: Measurements of large-scale live VoD from the 2008 Olympics. In *Proceedings of IMC* (2009).
11. Gulliver, S.R., Ghinea, G. Defining user perception of distributed multimedia quality. *ACM Trans. Multimed. Comput. Comm. Appl.* 2, 4 (Nov. 2006).
12. Yu, H., Zheng, D., Zhao, B.Y., Zheng, W. Understanding user behavior in large-scale video-on-demand systems. In *Proceedings of Eurosys* (2006).
13. Hei, X., Liang, C., Liang, J., Liu, Y., Ross, K.W. A measurement study of a large-scale P2P IPTV system. *IEEE Trans. Multimed.* 9 (2007).
14. Huang, Y., Tom, D.M.C., Fu, Z.J., Lui, J.C.S., Huang, C. Challenges, design and analysis of a large-scale P2P-VoD system. In *Proceedings of SIGCOMM* (2008).
15. Cho, K., Fukuda, K., Esaki, H. The impact and implications of the growth in residential user-to-user traffic. In *Proceedings of SIGCOMM* (2006).
16. Mitchell, T. *Machine Learning*, McGraw-Hill, 1997.
17. Saroiu, S., Gummadi, K.P., Dunn, R.J., Gribble, S.D., Levy, H.M. An analysis of Internet content delivery systems. In *Proceedings of OSDI* (2002).
18. Simon, H.A. Designing organizations for an information-rich world. *Computers, Communication, and the Public Interest*. M. Greenberger, ed. The Johns Hopkins Press.
19. Sripanidkulchai, K., Ganjam, A., Maggs, B., Zhang, H. The feasibility of supporting large-scale live streaming applications with dynamic application end-points. In *Proceedings of SIGCOMM* (2004).

Florin Dobrian, Asad Awan, Dilip Joseph, Aditya Ganjam, and Jibin Zhan Conviva, San Mateo, CA.

Vyas Sekar Stony Brook University, Stony Brook, NY.

Ion Stoica University of California, Berkeley Conviva, San Mateo, CA.

Hui Zhang Carnegie Mellon University, Pittsburgh, PA, Conviva, San Mateo, CA.

The Ultimate Online Resource for Computing Professionals & Students

ACM DL DIGITAL LIBRARY

<http://www.acm.org/dl>



Association for
Computing Machinery

Advancing Computing as a Science & Profession

Dalhousie University

Halifax, Canada

Faculty of Computer Science

Probationary Tenure Track Assistant Professor

Probationary Tenure Track Assistant Professor position in the Faculty of Computer Science. Dalhousie University (<http://www.dal.ca>) invites applications for a Probationary Tenure Track position at the Assistant Professor level in the Faculty of Computer Science (<http://www.cs.dal.ca>) that currently has 30 faculty members, approximately 425 undergraduate majors and 240 master's and doctoral students. The Faculty partners with other Faculties in the University to offer the Master of Electronic Commerce, Master of Health Informatics and Master of Science in Bioinformatics programs, and is an active participant in the Interdisciplinary PhD program. Dalhousie University is located in Halifax, Nova Scotia (<http://www.halifaxinfo.com/>), which is the largest city in Atlantic Canada and affords its residents outstanding quality of life.

The Faculty welcomes applications from outstanding candidates in Computer Science. An applicant should have a PhD in Computer Science or related area and be comfortable teaching core computer science courses, particularly Software Engineering. Evidence of a strong commitment

to and aptitude for research and teaching is essential. The ideal candidate will be open to collaborative research within the faculty and add to or complement existing research strengths and strategic research directions of the Faculty.

Applications should include an application letter, curriculum vitae, a statement of research and teaching interests, sample publications, and the names, email addresses and physical addresses of three referees. The application must include the Equity Self-Identification form (see the URL below). All documents are to be submitted to the email address below as PDF files.

Applicants should provide their referees with the URL of this advertisement (see below), and request that they forward letters of reference by email to the same address.

Applications will be accepted until April 30, 2013

All qualified candidates are encouraged to apply; however Canadian and permanent residents will be given priority. Dalhousie University is an Employment Equity/Affirmative Action Employer. The University encourages applications from qualified Aboriginal people, persons with a disability, racially visible persons and women.

Submission Address for application documents and reference letters:

appointments@cs.dal.ca

Location of this advertisement:

www.cs.dal.ca

Self-Identification form (PDF):

http://hrehp.dal.ca/Files/Academic_Hiring_%28For/selfid02.doc

Self-Identification form (Word):

http://hrehp.dal.ca/Files/Academic_Hiring_%28For/selfid02.doc

Iowa State University

Chair & Professor, Department of Computer Science

Iowa State University is seeking a Chair for the department of Computer Science. Primary responsibilities are to provide visionary leadership; encourage all aspects of research, teaching and service; advance development of department; facilitate faculty efforts to attract extramural funding; and promote productive relationships with all constituents. Successful applicants will have excellent communication skills.

ISU is an Affirmative Action/Equal Opportunity Employer. Please apply at www.iastatejobs.com, #121311.

Middle East Technical University Northern Cyprus Campus (METU NCC) Faculty Positions at Computer Engineering Program

Full-time faculty position (open rank) in Computer Engineering at METU NCC from September, 2013 (or an agreed date). Ph.D. in Computer Science or related field required, visiting positions available. Competitive salaries and subsidized accommodation offered, visit ncc.metu.edu.tr/academic/Guidelines_for_Application.php.



ADVERTISING IN CAREER OPPORTUNITIES

How to Submit a Classified Line Ad: Send an e-mail to acmm mediasales@acm.org. Please include text, and indicate the issue/ or issues where the ad will appear, and a contact name and number.

Estimates: An insertion order will then be e-mailed back to you. The ad will be typeset according to CACM guidelines. NO PROOFS can be sent. Classified line ads are NOT commissionable.

Rates: \$325.00 for six lines of text, 40 characters per line. \$32.50 for each additional line after the first six. The MINIMUM is six lines.

Deadlines: 20th of the month/2 months prior to issue date. For latest deadline info, please contact:

acmm mediasales@acm.org

Career Opportunities Online: Classified and recruitment display ads receive a free duplicate listing on our website at:

<http://jobs.acm.org>

Ads are listed for a period of 30 days.

For More Information Contact:

ACM Media Sales
at 212-626-0686 or
acmm mediasales@acm.org

The banner features the Inroads logo with the tagline "Paving the way toward excellence in computing education". It includes the URL <http://inroads.acm.org> and the ACM logo. Below the logo, it says "Association for Computing Machinery" and "Advancing Computing as a Science & Profession". The main text reads "ACM Inroads" and "The magazine for computing educators worldwide". At the bottom, it says "Paving the way toward excellence in computing education".



Peter Winkler

DOI:10.1145/2428556.2428578

Puzzled

Solutions and Sources

Last month (February 2013) we posed a trio of brainteasers concerning probability and dice. Here, we offer solutions to all three. How did you do?

1. Four different numbers.

You were asked to compute the probability of getting exactly four different numbers when tossing six dice (each with faces numbered 1 through 6). The answer, surprisingly (to me, at least, when I discovered it by accident) is that it is over 50%; that is, rolling a “four” this way is more likely than all other outcomes combined. The calculation is easy to mess up, by over- or undercounting. Four different numbers can be produced in two general ways: three of one number, and one each of three other numbers; or two each of two numbers, and one each of two others. For each of these ways, we first pick a “pattern” (such as ABCBBD), then assign numbers left to right. The total number of ways to roll a “four” is

$$((6 \text{ choose } 3) + (6 \text{ choose } 2) \times (4 \text{ choose } 2) \times 1/2) \times 6 \times 5 \times 4 \times 3 = 23,400$$

which, when divided by the total of 6^6 ways to roll the dice, gives approximately 50.154321%.

2. Four of a kind vs. six different.

This and the next puzzle were both inspired by correspondence with my longtime friend Bob Henderson of Mason, MI. A single die is rolled until Alice sees all the faces or Bob sees

four of a kind. One of them must win by the 16th roll; in the slowest case, 15 rolls would consist of three of each of five faces, and none of the sixth. Who is more likely to win? It turns out Bob has a sizeable advantage. The easiest way to see this is to prove by induction that after any number of rolls, he is more likely to have won than Alice. Surprised? I often see eyebrows go up at the idea that most of the time (over 63%, it turns out) some number will appear four times before some other number appears even once. This outcome is related to the oft-cited observation that runs of heads or tails in a sequence of coin flips tend to be longer than people expect.

3. Seven-seven vs. eight-seven.

A pair of dice is rolled repeatedly, with Alice gunning for two 7s in a row and Bob for an 8 followed immediately by a 7. Which is smaller: average time for Alice to succeed or average time for Bob to succeed? And who is more likely to get their wish first? Suppose the dice-pair is rolled twice. Of the 36^2 possible outcomes, the number yielding a sum of seven twice is $6^2 = 36$. The number of outcomes yielding 8 then 7 is only $5 \times 6 = 30$. Looks good for Alice, one might think. It takes on average six rolls to get a 7 and a seventh roll to try to duplicate it; Alice

will go through this procedure an average of six times, so, overall, it takes an average of $7 \times 6 = 42$ rolls for her to get what she wants. (We are making heavy use here of the fact that if you try something repeatedly that has probability of success p , it will take you on average $1/p$ trials to succeed.) The calculation for Bob is trickier since if he rolls an 8 then misses by rolling a second 8, he immediately gets another try. Doing the math, it works out to an average of 43.2 rolls to satisfy Bob. But another calculation shows that when Alice and Bob compete head to head on the same rolls, Bob wins with probability $35/66 > 53\%$. But how can this be? Perhaps the best intuition is provided by rolling the dice until *both* parties succeed. When Alice wins (that is, 7-7 comes up before 8-7) it is always by at least two rolls. But Bob can win by just one roll, when 8-7-7 comes up in that order. It is in just this kind of situation—where wins by one party tend to exhibit smaller margins than wins by the other—that expectation and probability can produce opposing results.

All readers are encouraged to submit prospective puzzles for future columns to puzzled@cacm.acm.org.

Peter Winkler (puzzled@cacm.acm.org) is William Morrill Professor of Mathematics and Computer Science, at Dartmouth College, Hanover, NH.

© 2013 ACM 0001-0782/13/03

[CONTINUED FROM P. 104] problems in the presence of failures in synchronous models, in the form of Byzantine agreement. They were also studying fault-tolerant clock synchronization. From that problem, I defined an easier problem of “approximate agreement” on real values, where everybody starts from a real value and has to agree on some value that is in the range of all the other values. We studied that first in synchronous models, and then we saw we could extend the result to asynchronous models. Putting it all together, it seemed pretty natural to consider the problem of exact agreement in asynchronous systems.

Another impetus was the then-current work on database transaction commits. This is a critical example of a practical problem of exact agreement on whether a transaction should commit or abort. It is important in practice for the solution to tolerate some failures, though not necessarily Byzantine failures—just simple stopping failures. And an asynchronous model would be appropriate, because you couldn’t realistically assume absolute bounds on the message delays.

How did your work proceed from there?

At first I thought that we might come up with an algorithm for the asynchronous case of this problem, like we had for approximate agreement. But our attempts failed, so we started trying to find an impossibility result. We went back and forth, working on both directions. We narrowed in on the solution relatively quickly—it didn’t take more than a few weeks. Formulating the ideas nicely, in terms of concepts like bivalence, came a bit later.

When did you realize FLP’s significance?

I think we understood the practical significance for transactions relatively quickly, but we did not predict the impact it would have on later research. Theoreticians have developed many results that extend FLP to other problems, and many results that circumvent the limitation using such methods as randomization and failure detectors. Most interestingly, I think, is that FLP triggered the development of algorithms that established a clear separation of requirements for fault-tolerant consensus problems: safety

properties of agreement and validity, which are required to hold always, and termination properties, which are required to hold during stable periods. These algorithms are not only interesting theoretically, but provide interesting guidelines for development of practical fault-tolerant systems.

In the 1980s, you also began work on input-output, or I/O, automata, which are used to model distributed algorithms.

Mark Tuttle and I developed the I/O automata modeling framework for asynchronous distributed systems early on, in 1987. We had some asynchronous distributed algorithms and we wanted to prove that they worked, but we were doing a lot of work to define our models and found that we were repeating that work in different papers. So we stepped back and developed a rigorous math model for systems with interacting components.

Later, you extended the work to cover synchronous systems, as well.

The I/O automata framework doesn’t deal with timing, so we defined another model, the Timed I/O Automata model, to cover synchronization. This is what we use as the foundation of our work on algorithms for mobile systems and wireless networks. My student Roberto Segala also worked with me to develop probabilistic versions, which are useful for describing randomized algorithms and security protocols.

So you have various frameworks that support the description of individual components in a system, and can then be used to produce a model for the entire system.

I don’t think the effort in developing these models is done yet. It would be nice to combine all the frameworks into one that includes discrete, continuous, timed, and probabilistic features, which is what’s needed to understand modern systems.

Let’s talk about some of your more recent work.

For the past 10 years or so, my group and I have been working on distributed algorithms for dynamic networks, in which the network changes over time because participating nodes can join, leave, fail, recover, and move, all while

the algorithm is operating. We have designed algorithms that maintain consistent data, synchronize clocks, compute functions, and coordinate robots. We have also worked quite a bit recently on low-level wireless communication issues—managing contention among different senders in wireless networks.

Are there certain techniques, principles, or characteristics you have found helpful, or does every fickle network bring its own set of problems?

Some common techniques emerge. For example, we try to implement abstraction layers, which are basically simpler models, over more complex models. You could have a Virtual Node layer that adds fixed nodes at known locations to a mobile wireless network and makes it easier to write higher-level algorithms. Or a Reliable Local Broadcast layer that masks issues of contention management in wireless networks, producing a more reliable substrate for writing higher-level algorithms.

Various algorithmic techniques do also recur, such as quorum-based reliable data management, random methods for information dissemination, and back-off protocols for scheduling transmissions.

You have also begun working on biologically inspired distributed algorithms. Can you talk a bit about that work?

I’m really just starting on this, but the idea is that biological systems are a lot like distributed algorithms. Why? Because they consist of many components, interacting to accomplish a common task, and communicating mainly with nearby components. So I’m reading about, for example, self-organizing systems of insects and bacteria, systems of cells during development, and neural networks, and I’m trying to apply a distributed algorithms viewpoint. It’s too early to see what will emerge, but surely we can define models, state problems, describe systems at different levels of abstraction as distributed algorithms, analyze the algorithms, and maybe even prove lower bounds.

Leah Hoffmann is a technology writer based in Brooklyn, NY.

© 2013 ACM 0001-0782/13/03

DOI:10.1145/2428556.2428579

Leah Hoffmann

Q&A

The Power of Distribution

Nancy Lynch talks about achieving consensus, developing algorithms, and mimicking biology in distributed systems.

DRAWN TO THE subject by its elegance, MIT professor Nancy Lynch has spent her career making sense of computational complexity while establishing the theoretical foundations of distributed computing. The FLP impossibility proof, among her best-known results, helped define the limitations of distributed systems. Input-output automata offered a valuable framework for verifying distributed algorithms. More recently, she has helped develop algorithms for dynamic networks and, during a fellowship year at the Radcliffe Institute for Advanced Study, begun to investigate a distributed approach to biological systems.

You were born into modest circumstances in Borough Park, Brooklyn. What drew you to math and computer science?

I don't come from an academic family. But I did well in math and got into Hunter High School, which was, at the time, a school for gifted girls—it is now co-ed. At Hunter, I had a wonderful mentor, Dr. Harry Ruderman, who adopted me as his protégé and encouraged me to explore advanced math problems. Then I went to Brooklyn College, took the Putnam exam, and got a great deal of attention and encouragement from the math faculty because I ranked in the top 80 or so nationwide. And after all that, I got into MIT with an NSF graduate fellowship.

Was it at MIT you were introduced to the field of theoretical computer science?



Right away, I lucked into taking Hartley Roger's course on recursive functions. I also took Seymour Papert's course on automata theory. I took other classes, of course—traditional math courses like algebra and analysis—but when it came time to choosing a research project, it seemed like all the other topics were already very well developed, and that it would be hard to make a big contribution. So at that point, two years in, I moved toward the newer areas of computational complexity theory and algorithms, where there was much more opportunity to have an impact. I was lucky enough to join a new and active group working in these areas, led by Albert Meyer and Mike Fischer.

After you finished your Ph.D., you had a series of jobs in the math departments of Tufts, the University of Southern California, and Florida International.

Yes, my husband and I had a two-body problem, so we kind of moved around. But math departments were not hiring very much, so in 1977, I went to Georgia Tech as an associate professor of computer science. At Georgia Tech, I was surrounded by applied computer scientists, so I abandoned working in abstract complexity theory and started looking at computer systems. Distributed systems were just beginning to be important at that time, and there were other people at Georgia Tech who were interested in building them. I decided there must be some interesting mathematics to be developed, and began to work on developing a theory for distributed systems.

That work put you back in touch with Michael Fischer, with whom you had worked at MIT.

Yes, Mike and I started working together on this, going back and forth between Georgia Tech and the University of Washington, where he was at the time. We made a lot of progress on this new theory very quickly, and in 1981, on the strength of that work, I went to MIT on a sabbatical, got a tenured offer the next year, and stayed. And I have been here ever since.

Your most famous result in distributed computing is the so-called FLP impossibility proof of 1985, which proves that asynchronous systems cannot reach consensus in the presence of one or more failures. Can you talk about how you reached it?

We were studying different models of distributed computing, both synchronous and asynchronous. In synchronous models, computation occurs in lock-step rounds. In asynchronous models, there is no common notion of time, and processes can move at arbitrarily different speeds.

Researchers like Leslie Lamport, Danny Dolev, and Ray Strong were studying consensus [CONTINUED ON P. 103]

Computing Reviews presents

The Best Reviews and
Notable Books & Articles
of 2012

Coming in April
online and in print

computingreviews.com

A daily snapshot of what is new and hot in computing.

4th Annual ACM SIGPLAN Conference on
Systems,
Programming,
Languages,
Applications:
Software for
Humanity

Submission Deadlines

March 28, 2013

- OOPSLA Papers
- Wavefront Papers & Experience Reports
- Proposals for Workshops & Panels

April 5, 2013

- Onward! Papers & Essays

June 8, 2013

- Dynamic Languages Symposium

June 28, 2013

- Posters, Doctoral Symposium
- ACM Student Research Competition
- Demonstrations
- Student Volunteers

Location

Hyatt Regency Indianapolis

Events

- 28th Annual OOPSLA
- Onward!
- Wavefront
- Dynamic Languages Symposium (DLS)
- Generative Programming & Component Engineering (GPCE)
- Software Language Engineering (SLE)
- ...and more

General Chairs

Patrick Eugster & Antony Hosking
Purdue University

OOPSLA Papers Chair

Cristina Lopes
University of California, Irvine

Onward! Papers Chair

Robert Hirschfeld
Hasso-Plattner-Institut Potsdam

Onward! Essays Chair

Bernd Brügge
Technische Universität München

DLS Papers Chair

Carl Friedrich Bolz
Heinrich-Heine-Universität Düsseldorf

More Information

<http://splashcon.org>

info@splashcon.org



SPLASH
INDIANAPOLIS 2013
OCTOBER 26-31

