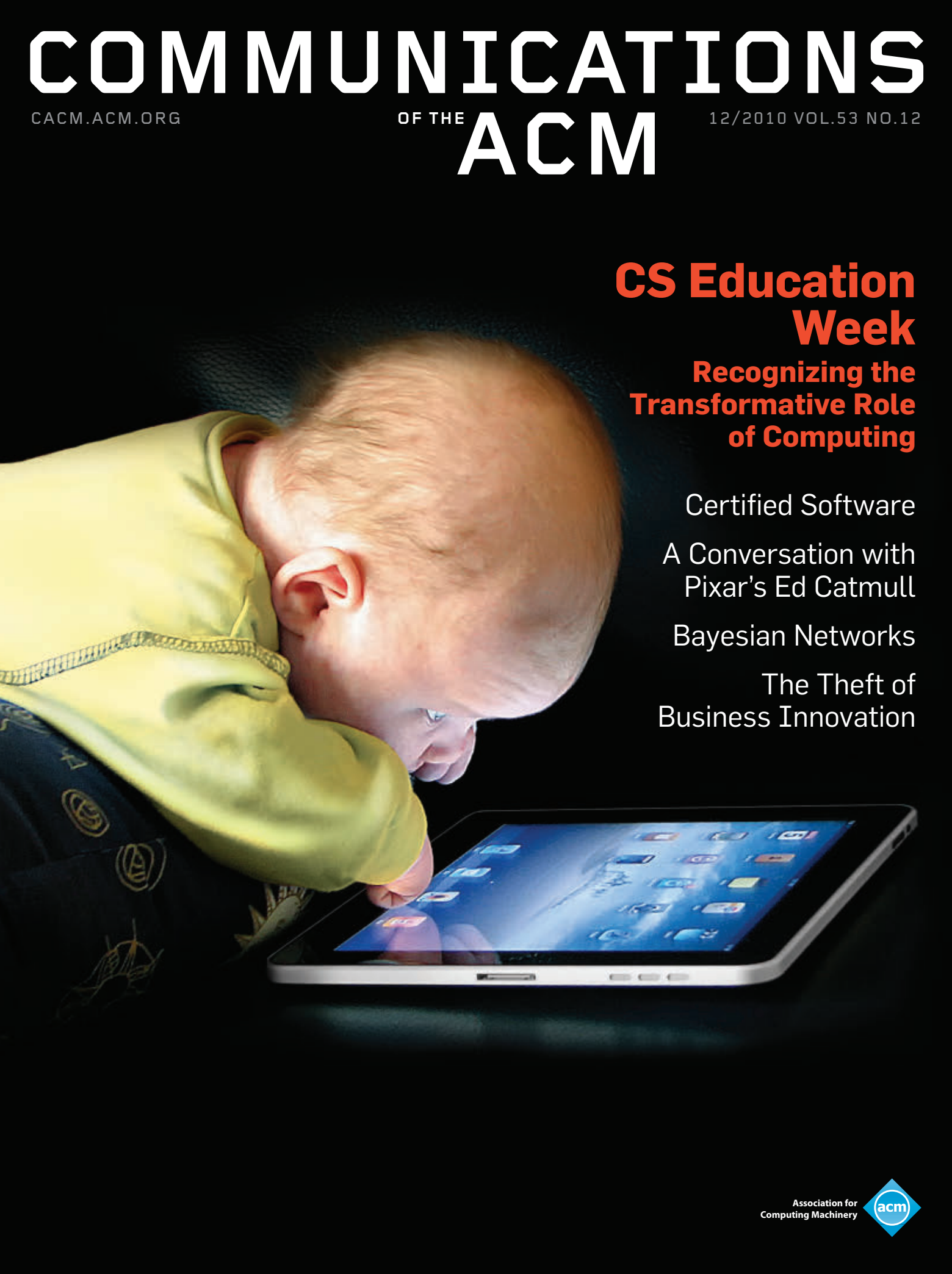
COMMUNICATIONS OF THE ACM

CACM.ACM.ORG

12/2010 VOL.53 NO.12



CS Education Week

Recognizing the Transformative Role of Computing

Certified Software

A Conversation with Pixar's Ed Catmull

Bayesian Networks

The Theft of Business Innovation

Association for Computing Machinery

vmware®

VMware Sponsored Academic Research Awards Request for Proposals (RFP)

Theme: Performance Management Challenges in Virtualized Environments

Virtualized Environments are proliferating in the IT industry. They are becoming a foundation of many computing and communication environments from large enterprises and multi-tenant clouds to virtualized desktops, as well as mobile endpoints. The management of performance in such virtualized environments provides many interesting challenges.

In this RFP, we are requesting proposals that cover but are not limited to any of the following areas:

- Development of large-scale statistics gathering and analysis in scalable virtualized environments. Specific areas of interest include but are not limited to health models for multi-tier applications as well as correlations application, VM and host performance.
- Improvements in coordinated resource management across applications, VMs and hosts. For example, transparent solutions for the double-swapping problem and elimination of redundant disk I/O across VM and host, as well as better management of runtime systems such as JVMs and databases in overcommitted situations in VM and host.
- Performance improvements of an emerging class of distributed, latency-sensitive programming and middleware frameworks, such as Hadoop, Memcached, GemFire Data Fabric and traditional HPC. This includes but is not limited to performance studies, performance optimizations, virtualization enhancements, and explorations of novel approaches that increase the value of these new frameworks through integration with virtualization.
- Performant and scalable handling of all virtualized environment management data, including data consistency, data distribution and levels of coupling between management and managed elements. For example, design of a scalable management and monitoring infrastructure for millions of VMs.

Award Information

Funding Award	up to $250,000
Duration of Proposed Research	One to three years (preference for 1 year projects, renewable based on results)
Funding Type	Gift or Grant (to be determined on a case by case basis)
Student Internships	One summer internship for each summer for each student for the duration of the project
Submission and Correspondence	funding@vmware.com
Program URL	http://vmware.com/go/vmap-rfp

Proposal Evaluation

Phase 1

Preliminary Proposal Page Length	One page short technical proposal with coarse grain deliverables
Preliminary Proposal Due Date	January 10, 2011
Preliminary Proposal Notification	February 4, 2011

Phase 2

***Optional* Exploratory Virtual Workshop**	Principal Investigators chosen for Phase 2 are invited to an optional ½ day virtual workshop on Performance Management Challenges in Virtualized Environments
Workshop Date	February 18, 2011
Full Proposal Page Length	10-20 pages
Full Proposal Content	Technical approach, preliminary results, deliverables, milestones, budget, CVs
Full Proposal Deadline	April 8, 2011
Award Results Announcement	April 29, 2011
Project Start	May 27, 2011

Springer
the language of science

springer.com

Springer References & Key Library Titles

Handbook of Ambient Intelligence and Smart Environments

H. Nakashima, Future University, Hakodate, Hokkaido, Japan; **H. Aghajan**, Stanford University, Stanford, CA, USA; **J. C. Augusto**, University of Ulster at Jordanstown, Newtownabbey, UK (Eds.)

Provides readers with comprehensive, up-to-date coverage in this emerging field. Organizes all major concepts, theories, methodologies, trends and challenges into a coherent, unified repository. Covers a wide range of applications relevant to both ambient intelligence and smart environments. Examines case studies of recent major projects to present the reader with a global perspective of actual developments.

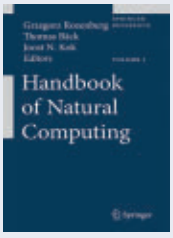
2010. XVIII, 1294 p. 100 illus. Hardcover
ISBN 978-0-387-93807-3 ► **$229.00**

Handbook of Multimedia for Digital Entertainment and Arts

B. Furht, Florida Atlantic University, Boca Raton, FL, USA (Ed.)

The first comprehensive handbook to cover recent research and technical trends in the field of digital entertainment and art. Includes an outline for future research directions within this explosive field. The main focus targets interactive and online games, edutainment, e-performance, personal broadcasting, innovative technologies for digital arts, digital visual and other advanced topics.

2009. XVI, 769 p. 300 illus., 150 in color. Hardcover
ISBN 978-0-387-89023-4 ► **$199.00**



Handbook of Natural Computing

G. Rozenberg, **T. Bäck**, **J. N. Kok**, Leiden University, The Netherlands (Eds.)

We are now witnessing an exciting interaction between computer science and the natural sciences. Natural Computing is an important catalyst for this interaction, and this handbook is a major record of this important development.

2011. Approx. 1700 p. (In 3 volumes, not available seperately) Hardcover
ISBN 978-3-540-92909-3 ► **$749.00**

eReference
ISBN 978-3-540-92910-9 ► **$749.00**

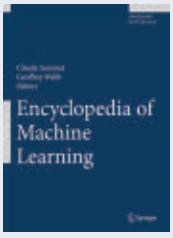
Print + eReference
2011. Approx. 1700 p.
ISBN 978-3-540-92911-6 ► **$939.00**

Handbook of Biomedical Imaging

N. Paragios, École Centrale de Paris, France; **J. Duncan**, Yale University, USA; **N. Ayache**, INRIA, France (Eds.)

This book offers a unique guide to the entire chain of biomedical imaging, explaining how image formation is done, and how the most appropriate algorithms are used to address demands and diagnoses. It is an exceptional tool for radiologists, research scientists, senior undergraduate and graduate students in health sciences and engineering, and university professors.

2010. Approx. 590 p. Hardcover
ISBN 978-0-387-09748-0 ► **approx. $179.00**



Encyclopedia of Machine Learning

C. Sammut, **G. I. Webb** (Eds.)

The first reference work on Machine Learning Comprehensive A-Z coverage of this complex subject area makes this work easily accessible to professionals and researchers in all fields who are interested in a particular aspect of Machine LearningTargeted literature references provide additional value for researchers looking to study a topic in more detail.

2010. 800 p. Hardcover
ISBN 978-0-387-30768-8 ► **approx. $549.00**

eReference
2010. 800 p.
ISBN 978-0-387-30164-8 ► **approx. $549.00**

Print + eReference
2010. 800 p.
ISBN 978-0-387-34558-1 ► **approx. $689.00**

Handbook of Peer-to-Peer Networking

X. Shen, University of Waterloo, ON, Canada; **H. Yu**, Huawei Technologies, Bridgewater, NJ, USA; **J. Buford**, Avaya Labs Research, Basking Ridge, NJ, USA; **M. Akon**, University of Waterloo, ON, Canada (Eds.)

Offers elaborate discussions on fundamentals of peer-to-peer computing model, networks and applications. Provides a comprehensive study on recent advancements, crucial design choices, open problems, and possible solution strategies. Written by a team of leading international researchers and professionals.

2010. XLVIII, 1500 p. Hardcover
ISBN 978-0-387-09750-3 ► **$249.00**

Easy Ways to Order for the Americas ► **Write:** Springer Order Department, PO Box 2485, Secaucus, NJ 07096-2485, USA ► **Call: (toll free)** 1-800-SPRINGER ► **Fax:** 1-201-348-4505 ► **Email:** orders-ny@springer.com or **for outside the Americas** ► **Write:** Springer Customer Service Center GmbH, Haberstrasse 7, 69126 Heidelberg, Germany ► **Call:** +49 (0) 6221-345-4301 ► **Fax :** +49 (0) 6221-345-4229 ► **Email:** orders-hd-individuals@springer.com
► Prices are subject to change without notice. All prices are net prices.

014712x

COMMUNICATIONS OF THE ACM

Departments

Last Byte

Association for Computing Machinery
Advancing Computing as a Science & Profession

News

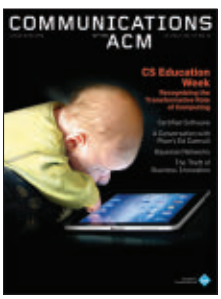
Viewpoints

PHOTOGRAPH BY JALANI MORGAN

12/2010
VOL. 53 NO. 12

Practice

Articles' development led by acmqueue
queue.acm.org

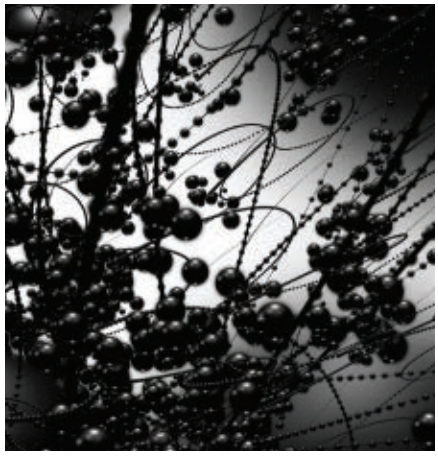


About the Cover: Computer Science Education Week, designated by the U.S. House of Representatives, recognizes the transformative role of computing and the need to bolster computer science at all educational levels. ACM will promote CSEdWeek (Dec. 5–11) to help raise awareness of computing and its role in preparing citizens of all ages to grow and prosper in the 21st century. Photograph by Steve Paine/Carrypad.com.

PHOTOGRAPH BY TOM UPTON; IMAGE BY ROBERT HODGIN

Contributed Articles

Review Articles



Research Highlights

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.

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.

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:10.1145/1859204.1859205

David A. Patterson

Diverse Connections

Which computer science conference: ▸ Includes a past President of the National Academy of Engineering, a past chair of IBM Academy of Technology, a past

president of ACM, a past chair of the Computing Research Association, several members of the National Academy of Engineering, and the current executive vice president of Research and Engineering at Google?

▸ Has Google, Intel, Microsoft, Cisco, and NetApp as major sponsors?

▸ Has attendance that averages 50% female, 40% African American, and 30% Hispanic?

The answer is the Richard Tapia Celebration of Diversity in Computing Conference, to be held April 3–5, 2011 (http://tapiaconference.org/2011/).

Created as a supportive networking environment for underrepresented groups in computing and information technology, the Tapia conference celebrates and attracts students and professionals from diverse backgrounds pursuing careers in our field. This biannual event features inspiring speakers, a dynamic technical program, and a community of encouragement and motivation. While there have long been computing conferences focused on individual minority groups, Tapia was founded as an *inclusive* conference transcending demographic changes over time. And it works!

A survey from Tapia 2009 found that 82% of the attendees agreed the conference increased their dedication to complete their degree and reaffirmed their belief that computing was the right career path for them. One reason is that the Tapia Conference helps people from underrepresented groups overcome the feelings of isolation sadly all too common in our field.

Tapia 2011 marks the 10th anniversary of the conference. Tapped to serve as general chair is David Patterson, UC Berkeley—past president of ACM, past chair of the CRA, member of the National Academy of Engineering and the National Academy of Sciences.

Some of the highlights planned for the event include:

▸ The Memorial Ken Kennedy Lecture will be given by Irving Wladawsky-Berger, former chair of the IBM Academy of Engineering and the HENAAC Hispanic Engineer of the Year;

▸ A speech on the future of IT by Alan Eustace, executive vice president for Research and Engineering at Google;

▸ Presentation of the 2011 Richard Tapia Award to William Wulf, University of Virginia, and past president of the National Academy of Engineering.

The program will feature luminaries like Deborah Estrin of UCLA, Blaise Aguera y Arcas of Microsoft, John Kubiatowicz of UC Berkeley, and rising stars like Ayanna Howard of Georgia Tech, Ilya Hicks of Rice University, and Patty Lopez of Intel. There will be workshops on resumé writing, professional development, and a doctoral consortium.

New this year is a focus on connections: between speakers and audience, among students and professionals, and beyond the conference via remote research collaborations. The program provides many opportunities to network and build relationships, including an opening reception, a poster session displaying exciting research by students, and a "meetup" day designed to match students with opportunities.

How Can You Help?

We predict this Tapia conference will be the largest yet. If you would like to help, here are some suggestions:

Faculty

▸ Encourage your undergraduate and graduate students to apply for a travel scholarship to Tapia; applications are available at http://tapiaconference.org/2011/scholarships.html;

▸ Provide departmental funds (or raise funds from industry) to send people not funded by scholarships.

▸ If your university sends at least 10 students, you can create a poster to help recruit students to your institution.

▸ If you are passionate about these topics, consider registering.

Professionals

▸ Encourage your staff to register at https://apply2.cse.tamu.edu/gts/applicant/TapiaConference/

▸ If your company would like to be a Tapia sponsor, visit http://tapiaconference.org/2011/supporters.html and contact Cynthia Lanius (funding@tapiaconference.org).

Students

▸ For those working on a Ph.D. dissertation, consider applying to present at the Doctorial Consortium. This event provides a highly technical and supportive environment where Ph.D. students present their research and receive feedback from a panel of experts; http://tapiaconference.org/2011/dc.html.

▸ If you have research results, apply to present a poster at http://tapiaconference.org/2011/posters.html.

▸ Apply for a scholarship at http://tapiaconference.org/2011/scholarships/apply.

Tapia 2011 promises to be inspirational, educational, and a lot of fun. We look forward to seeing you April 3 in San Francisco.

David A. Patterson (pattrsn@EECS.Berkeley.EDU).

DOI:10.1145/1859204.1859206

Science Has *Four* Legs

As an editor of *The Fourth Paradigm* (http://research.microsoft.com/en-us/collaboration/fourthparadigm/default.aspx, Microsoft Research, Redmond, WA, 2009) and someone who subscribes to Jim Gray's vision that there are now four fundamental scientific methodologies, I feel I must respond to Moshe Y. Vardi's Editor's Letter "Science Has Only Two Legs" (Sept. 2010).

First, I should explain my qualifications for defending the science-has-four-legs premise. From 1964, beginning as a physics undergraduate at Oxford, until 1984, when I moved from physics to the Electronics and Computer Science Department, I was a working natural scientist. My Ph.D. is in theoretical particle physics, and, in my research career, I worked extensively with experimentalists and spent two years at the CERN accelerator laboratory in Geneva. In computer science, my research takes in all aspects of parallel computing—architectures, languages, and tools, as well as methodologies for parallelizing scientific applications—and more recently the multi-core challenge. From 2001 to 2005, before I joined Microsoft, I was Director of the U.K.'s eScience Core Program, working closely with scientists of all descriptions, from astronomers and biologists to chemists and environmental scientists. Here at Microsoft Research, I still work with practicing scientists.

I therefore have some relevant experience on which to ground my argument. By contrast, though Vardi has had a distinguished career in mathematics and computer science (and has done a great job with *Communications*), he has not, as far as I know, had much direct involvement with the natural sciences.

It is quite clear that the two new scientific paradigms—computational and data-intensive—do not displace experiment and theory, which remain as relevant as ever. However, over the past 50 years it is equally clear that computational science has emerged as a third methodology with which we now explore problems that are simply inaccessible to experiment. To do so, scientists need (along with their knowledge of experiment and theory) training in numerical methods, computer architecture, and parallel programming. It was for this reason that Physics Nobel Prize laureate Ken Wilson in 1987 called computational science the "third paradigm" for scientific discovery. He was investigating quantum chromo-dynamics, or QCD, describing the fundamental equations between quark and gluon fields behind the strong nuclear force. No analytic solution is possible for solving these equations, and the only option is to approximate the theory on a space-time lattice. Wilson pioneered this technique, using supercomputers to explore the predictions of QCD in the physical limit when the lattice spacing tends to zero. Other examples of such computational exploration, including galaxy formation and climate modeling, are not testable through experiment in the usual sense of the word.

In order to explore new techniques for storing, managing, manipulating, mining, and visualizing large data sets, Gray felt the explosive growth of scientific data posed profound challenges for computer science (see Fran Berman's "Got Data? A Guide to Data Preservation in the Information Age," *Communications*, Dec. 2008). He therefore spent much of his last years working with scientists who were, as he said, "drowning in data." Working with astronomers allowed him the luxury of experimentation, since, he said, their data had "absolutely no commercial value." Similarly, the genomic revolution is upon us, and biologists need powerful tools to help disentangle the effects of multiple genes on disease, develop new vaccines, and design effective drugs. In environmental science, data from large-scale sensor networks is beginning to complement satellite data, and we are seeing the emergence of a new field of environmental informatics. Scientists require significant help not only in managing and processing the raw data but also in designing better workflow tools that automatically capture the provenance involved in producing the data sets scientists actually work with.

On the other hand, "computational thinking" attempts to demonstrate the power of computer science ideas not just in science but also in many other aspects of daily life. However, despite its importance, this goal should not be confused with the emergence of the two new methodologies scientists now need to assist them in their understanding of nature.

Tony Hey, Seattle

Author's Response:

Hey and I are in violent agreement that science today is thoroughly computational. What I fail to see is why this requires it to sprout new legs. In fact, theory in science was mathematical way before it was computational. Does that make mathematics another leg of science?

Experimental science always relied on statistical analysis. Does that make statistics another leg of science? Science today relies on highly complex theoretical models, requiring analysis via computation, and experimental setups that yield massive amounts of data, also requiring analysis via computation. So science is thoroughly computational but still has only two legs—theory and experiment.

Moshe Y. Vardi, Editor-in-Chief

Let Patients Participate in Their Own Care

In his article "Computers in Patient Care: The Promise and the Challenge" (Sept. 2010), Stephen V. Cantrill, M.D., offered seven compelling arguments for integrating health information technology (HIT) into

clinical practice. However, he missed one that may ultimately surpass all others—making medical data meaningful (and available) to patients—so they can be more informed partners in their own care.

Dr. Cantrill was not alone in appreciating the value of HIT this way. Patient-facing electronic data presentation is consistently overlooked in academic, medical, industrial, and political discussions, likely because it's much more difficult to associate financial value with patient engagement than with measurable inefficiencies in medical practice.

Perhaps, too, computer scientists have not let patients take advantage of the growing volume of their own electronic medical data; allowing them only to, say, download and print their medical histories is important but insufficient. Medical data is (and probably should be) authored by and for practitioners, and is thus beyond the health literacy of most patients. But making medical data intuitive to patients—a problem that's part pedagogy, part translation, part infrastructure, and part design—requires a collaborative effort among researchers in human-computer interaction, natural language processing, visualization, databases, and security. The effort also represents a major opportunity for CS in terms of societal impact. Its omission is indicative of just how much remains to be done.

Dan Morris, Redmond, WA

Release the Code

About the software of science, Dennis McCafferty's news story (Oct. 2010) asked "Should Code Be Released?" In the case of climate science code, the Climate Code Foundation (http://climatecode.org/) answers with an emphatic yes. Rebuilding public trust in climate science and support for policy decisions require changes in the transparency and communication of the science. The Foundation works with climate scientists to encourage publication of all climate-science software.

In a *Nature* opinion piece "Publish Your Computer Code: It Is Good Enough" (Oct. 13, 2010, http://www.nature.com/news/2010/101013/full/467753a.html), I argued that there are powerful reasons to publish source code across all fields of science, and that software is an essential aspect of the scientific method despite failing to benefit from the system of competitive review that has driven science forward for the past 300 years. In the same way software is part of the scientific method, source code should be published as part of the method description.

As a reason for not publishing software, McCafferty quoted Alan T. DeKok, a former physicist, now CTO of Mancala Networks, saying it might be "blatantly wrong." Surely this is a reason, perhaps the main one, that software *should* be published—to expose errors. Science progresses by testing new ideas and rejecting those that are wrong.

I'd also like to point out a glaring red herring in McCafferty's story—the suggestion that a policy in this area could undermine a modern-day Manhattan Project. All design and method descriptions from that project were top secret for years, many to this day. Such secrecy would naturally apply to any science software of similar military importance.

Nick Barnes, Staines, U.K.

The Brain's Inner Computer Is Analog

I continue to be amazed by the simplistic approach pursued by computer scientists trying to understand how the brain functions. David Lindley's news article "Brains and Bytes" (Sept. 2010) came tantalizingly close to an epiphany but didn't quite express what to me is fundamentally wrong with most research in the field. There is an appreciation of the statistical nature of the brain's functioning at the microscopic, cellular level, a realization that complete predictability is not only not achievable but actually completely inappropriate.

Lindley referred to an event ("neural firing") as a binary process, despite being statistical in its occurrence. Lacking personal experience (so unfettered by knowledge), I claim this represents the fundamental obstacle to achieving a true understanding of how the brain works. A neuron firing or a synapse transmitting the result is neither binary nor random; rather, the shape and strength of the "signal" are critical in achieving understanding, and are, for the most part, ignored.

Many researchers seem precommitted to a view defined by digital processes coupled with statistical unpredictability. Time to return to the Dark Ages of computing when the brain's cellular components were not statistically imperfect digital devices. They were and are analog, a word Lindley left out of the article, even though some of his descriptions of the cellular functions cried out for such a characterization.

R. Gary Marquart, Austin, TX

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

Association for Computing Machinery

Advancing Computing as a Science & Profession

membership application & *digital library* order form

Priority Code: AD10

You can join ACM in several easy ways:

Online	Phone	Fax
http://www.acm.org/join	*+1-800-342-6626 (US & Canada)* *+1-212-626-0500 (Global)*	*+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/serving/ethics.html

choose one membership option:

PROFESSIONAL MEMBERSHIP:

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

STUDENT MEMBERSHIP:

- o **ACM Student Membership: $19 USD**
- o **ACM Student Membership plus the ACM Digital Library: $42 USD**
- o **ACM Student Membership PLUS Print *CACM* Magazine: $42 USD**
- o **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*. 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.

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

o Professional Member Dues ($99 or $198) $ _____

o ACM Digital Library ($99) $ _____

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

Total Amount Due $ _____

Card # Expiration date

Signature

DOI:10.1145/1859204.1859207

In the Virtual Extension

To ensure the timely publication of articles, Communications *created the Virtual Extension (VE) to expand the page limitations of the print edition by bringing readers the same high-quality articles in an online-only format. VE articles undergo the same rigorous review process as those in the print edition and are accepted for publication on merit. The following synopses are from articles now available in their entirety to ACM members via the Digital Library.*

viewpoint

DOI: 10.1145/1859204.1859234

The Role of Conference Publications in CS

Massimo Franceschet

The role of conference publications in computer science is controversial. Conferences have the undeniable advantages of providing fast and regular publication of papers and of bringing researchers together by offering the opportunity to present and discuss the paper with peers. These peculiar features of conferences are particularly important because computer science is a relatively young and fast-evolving discipline.

Recently, *Communications* published a series of thought-provoking Viewpoint columns and letters that swim against the tide. These contributions highlight many flaws of the conference system, in particular when compared to archival journals, and also suggest a game-based solution to scale the academic publication process to Internet scale. Some of the mentioned flaws are: short time for referees to review the papers, limited number of pages for publication, limited time for authors to polish the paper after receiving comments from reviewers, and overload of the best researchers as reviewers in conference program committees.

This article gives an alternative view on this hot issue: the bibliometric perspective. Bibliometrics has become a standard tool of science policy and research management in the last decades. In particular, academic institutions increasingly rely on bibliometric analysis for making decisions regarding hiring, promotion, tenure, and funding of scholars. This article investigates the frequency and impact of conference publications in computer science as compared with journal articles. The set of computer science publications is stratified by author, topic, and nation; in particular, publications of the most prolific, most popular, and most prestigious scholars in computer science are analyzed and evaluated.

contributed article

DOI: 10.1145/1859204.1859236

IT 2008: The History of a New Computing Discipline

Barry Lunt, J. Ekstrom, Han Reichgelt, Michael Bailey, and Richard LeBlanc

The early 1990s saw the emergence of the Internet from the environs of the technical cognoscenti into the dot-com world with an interface for the masses. The increased complexity and importance of computing technologies for the success of organizations and individuals led to a growing need for professionals to select, create, apply, integrate, and administer an organizational IT infrastructure. The skill sets needed for the new breed of network and system administrators were not provided by the computer science programs of the time. Moreover, information systems programs, with the business education requirements of their accreditation bodies, were equally unwilling or unable to include the technical depth required.

In response to this new educational need, university programs arose that were called Information Systems and Computer Science, respectively, but were something else entirely. These programs, and others like them, sprung up independently and spontaneously to satisfy the needs of employers for workers with skills in networks, distributed systems, and beginning in the mid-1990s, the Web.

On the national level, the Computing Sciences Accrediting Board was joining with ABET. Within ABET both the newly formed Computing Accreditation Commission and the Technology Accreditation Commission had noticed the emerging IT programs, and were wondering under which commission IT would best fit. It was in this lively environment that a group was formed that would guide IT through the period of defining its own model curriculum, its place with respect to the other computing programs already extant, and its own accreditation criteria.

contributed article

DOI: 10.1145/1859204.1859235

A Global Collaboration to Deploy Help to China

Ralph Morelli, Chamindra de Silva, Trishan de Lanerolle, Rebecca Curzon, and Xin Sheng Mao

On May 12, 2008, an earthquake measuring 7.9 on the Richter scale struck in Sichuan Province in southwestern China, destroying homes, schools, hospitals, roads, and vital power and communication infrastructure. More than 45 million people were affected—tens of thousands were killed, hundreds of thousands injured, millions of people were evacuated and left homeless, and millions of buildings were destroyed.

When the earthquake hit, several members of what became an international, volunteer, disaster-management IT team were attending a workshop in Washington, D.C. organized by IBM to train IBM personnel and others in the use and deployment of Sahana, a free and open source software (FOSS) disaster management system.

Sahana is a Web-based collaboration tool that helps manage information resources during a disaster recovery effort. It supports a wide range of relief efforts from finding missing persons, to managing volunteers, tracking resources, and coordinating refugee camps. Sahana enables government groups, non-governmental organizations (NGOs), and the victims themselves to work together during a disaster recovery effort.

This article provides a firsthand account of an international team effort to install the Sahana system in Chengdu, Sichuan. It describes how a diverse, multidisciplinary team worked together to assist the earthquake recovery effort. The success of the collaboration illustrates the power of virtual communities working across international boundaries using a variety of communication software. It also demonstrates that the Internet has truly made us all neighbors and is constantly forcing us to redefine our concept of community.

BLOG@CACM

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.

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

DOI:10.1145/1859204.1859208 http://cacm.acm.org/blogs/blog-cacm

Security Advice; Malvertisements; and CS Education in Qatar

Greg Linden discusses security advice and the cost of user effort, Jason Hong considers the increase in malvertisements, and Mark Guzdial writes about gender and CS education in Qatar.

Greg Linden
"What Security Advice Should We Give?"
http://cacm.acm.org/blogs/blog-cacm/87847

Should people follow the security advice we give them?

The surprising answer is no. According to a recent paper, "So Long, And No Thanks for the Externalities: The Rational Rejection of Security Advice by Users," by Cormac Herley at Microsoft Research, not only do people not follow the security advice we give them, but they shouldn't.

The problem is that security advice ignores the cost of user effort. When the likelihood of having a loss is low, and if the cost of the loss in time or money is low, then the cost of being vigilant must be trivially low. Much of what we ask of people takes too much effort. Taking an example from Herley's paper, if only 1% per year get hit with a threat that costs 10 hours to clean up, the effort required to avoid the threat must be no more than one second per day.

This is a frighteningly low bar. It means that almost all end-user security must require nearly no effort.

Can security features have this little effort?

Some do. For example, rather than imposing harsh and mandatory restrictions on passwords (for example, length between 6–8 characters, must contain a number and a letter, must be changed every three weeks), some Web sites merely report an estimate of the strength of a password while accepting almost anything. This imposes almost no effort while still encouraging longer, stronger, and more memorable passwords. Not only does this make sense for users, but it also makes sense for companies since, as Herley's paper points out, the costs of having more agent-assisted password resets after forcing people to choose difficult-to-remember passwords can easily be higher than the cost of having more attacks.

Another example implemented by some browsers is improving the visibility of the domain when displaying the URL in a browser. This makes it much easier to see if you are at the correct Web site, possibly reducing that effort below the threshold where people will find it worthwhile.

A third example is the anti-phishing feature now common in Web browsers. This feature checks if a Web site is a known security threat and intervenes in the rare cases where someone visits a known threat. The cost of this is zero for almost all Web browsing as the feature is working quietly behind the scene.

Perhaps the question at the beginning of this post is wrong. Perhaps we should ask not whether people should follow the security advice we give them, but what advice we should be giving. The security advice we give has to consider the cost of user effort. The security advice we give also has to be worth following.

So, what security advice should we be giving?

Jason Hong
"Malvertisements Growing as Online Security Threat"
http://cacm.acm.org/blogs/blog-cacm/90522

I'm at the Anti-Phishing Working Group's Counter eCrime Operations Summit IV this week. The conference is attended by law-enforcement officers, researchers, and industry professionals. I'll be giving some highlights that are relevant to usable privacy and security.

Gary Warner from University of Ala-

bama reported on trends in malvertising, a relatively new kind of attack where criminals inject malware or scareware into online advertisements. These malvertisements, for example, might be Flash files that make use of exploits, or use scare tactics that "warn" users about viruses that are on their computer and urge people to click a link to install fake antivirus software.

There are three points I want to discuss. First, these advertising networks have a very wide reach on the Internet. Even the *New York Times*' Web site was hit with one of these fake advertisements. As such, these malvertisements represent a very serious threat to the operation of the Internet.

Second, as a user, you could be doing everything right and *still* be infected. You might keep your antivirus software up to date, always install the latest patches, avoid sketchy programs and Web sites, and not fall for any phish, but still end up with malware.

Third, using fake virus scans has been a growing tactic to convince people to install malware onto their computers. This kind of malware is growing in sophistication, and is also causing damage to legitimate antivirus vendors by reducing people's trust. Admittedly, it's a good strategy for the bad guys to take.

I had the misfortune of facing some of this fake antivirus software recently. My wife fell for one of these scams and asked me to fix her computer. The malware actually blocked standard antivirus software from running, so I tried to remove the software manually. However, as I did this, I saw the malware start to reinstall itself from a remote location. I tried again after turning off all networking, and deleted all the malware files. However, I either missed something in the registry or a browser helper object as it started reinstalling itself again after rebooting. After wasting an hour of time, we decided it would be easier and safer to just wipe the machine and start over.

If we take a step back, we can view malvertisements as just another type of attack where criminals try to make use of our greater connectivity. It's useful to revisit the three basic strategies for usable privacy and security: 1) make it invisible; 2) provide better user interfaces; and 3) educate users. In the short term, we can educate people about fake antivirus programs. However, in the long term, advertising networks will need far better tools for detecting and filtering these kinds of malware so users don't see them at all.

Mark Guzdial "The Complicated Issues of Computing Education in Qatar"

http://cacm.acm.org/blogs/blog-cacm/91580

At the beginning of May, the ACM Education Board visited Qatar University. The goal was to meet with stakeholders and plan for developing computing education in the Middle East and India. John Impagliazzo, professor at Qatar University (QU), longtime Education Board member, and emeritus professor from Hofstra University, organized the meeting. We went with Dame Wendy Hall, ACM President [at the time], and John White, ACM CEO and executive director. The trip was amazing—enlightening and confounding.

Qatar University has a significant gender imbalance in its computer science program, but it's opposite of the U.S. and much of the Western world. Seventy percent of QU's CS students are female. The QU professors explained that being a computer scientist isn't a well-respected job in Qatar. The men go for engineering-labeled degrees, which lead to higher-paying jobs. The QU faculty explained that the computing jobs in Doha—the capital of Qatar, home of QU, and of the majority of Qatar's population—are mostly about adapting and customizing applications from elsewhere to make them fit in Qatar and the Middle Eastern culture.

The QU CS faculty are planning to add more information systems and information technology into their curriculum to better prepare their students for the available jobs. Then we got to meet the female CS students at QU. These women totally embrace "geekiness." They are *pushing* their faculty to let them *build* applications sooner in the curriculum. QU offers no programming competitions. These women are looking up international programming competition problems to push themselves! I was amazed at how eager these women are, how much they want to program "robots, animation, mobiles—anything! We want to be challenged!"

On the next day, we visited Education City, home to the satellite campuses of six American universities. Built by the Qatar Foundation, the goal is to change Qatari culture into a knowledge-based society where people build their own intellectual property and not just adapt Western ones. Qatar has oil wealth now, but knows it won't last forever. They're investing now for a future society with a culture focused on knowledge creation and innovation.

CMU Qatar (CMUQ) has a computer science program that is 50/50 female and male! They emphasize developing "a Geek culture," because they want to encourage the sense of wanting to learn and digging in to figure it out yourself, which they saw as missing from the local culture. Classes at CMUQ are coed, integrating men and women. Not all CMUQ students are from Doha, while most of the women at QU are from Doha, and have no interest in leaving Doha.

Why are the women of QU not going to CMUQ? They told us because they *want* segregated classes. They don't want to go to classes with men. QU has a women's campus and a men's campus. Many women at QU never go to the men's campus.

When we met with the CMUQ faculty, they told us they want to see *more* "computer science" curriculum in Doha, and less information systems and information technology. That's what the Qatar Foundation wants, to see more about creating new technologies, not adopting, adapting, and managing existing ones. Are there jobs for that in Doha? "Maybe not now, but there will be!"

But will the new jobs be there in time for these students who won't leave Doha? How quickly can culture (and industry) change to embrace innovation over adaptation? The CS women of Qatar University are *eager* to program, *hungry* to build new applications, but just as intensely *value* their gender segregation and *staying* with their families in Doha. It's a complicated and fascinating problem of the challenges of changing the culture of a nation.

Greg Linden is the founder of Geeky Ventures. **Jason Hong** is an associate professor of computer science at Carnegie Mellon University. **Mark Guzdial** is a professor at the Georgia Institute of Technology.

DOI:10.1145/1859204.1859209 David Roman

School Grades Need Improvement

"Technology has been paying the bills in this country...we're killing the goose that laid the golden eggs." —Stan Williams, Senior Fellow, Hewlett-Packard

Much has been written over the last decade about the abysmal state of the education arms race in the U.S., particularly in the areas of science, technology, engineering, and math (popularly known as STEM) with increasing low blows to computer science. But two recently published reports make it painfully clear just how little has been done to redirect that trend over the last 10 years and what is bound to happen if this spiral does not end.

ACM and the Computer Science Teachers Association (CSTA) recently released some startling findings of an in-depth study on U.S. CS education standards in a report entitled *Running on Empty: The Failure to Teach K–12 Computer Science in the Digital Age* (http://www.acm.org/runningonempty/). The report found that approximately two-thirds of U.S. states have very few CS education standards for secondary school education, and most states treat high school CS courses as simply an elective and not part of a student's core education. The report also discovered that only 14 states have adopted significant education standards for high school CS programs, and 14 states and the District of Columbia have not adopted *any* upper-level standards for CS instruction.

Secondary schools offering introductory (or pre-AP) Computer Science courses, change from 2005 baseline

2007	2009
–6%	–17%

Secondary offering AP Computer Science courses, change from 2005 baseline

2007	2009
–20%	–35%

Source: Computer Science Teachers Association survery data of high schools

The system's failures in the STEM disciplines also raise concerns about the U.S. status as a place to innovate, invest in the future, and create high-paying jobs, according to *Rising Above the Gathering Storm, Revisited*, a report published by the National Academy of Sciences (http://www.nap.edu/catalog.php?record_id=12999). This recent study looks at changes that have occurred since the 2005 publication of the landmark, 500-page report known as the "Gathering Storm," which focused upon the ability of Americans to compete for employment in a job market that "increasingly knows no geographic boundaries." *Revisited* opens with the telling comment: "Five years have passed since the initial report was prepared, a period in which a great deal has changed...and a great deal has not changed." It goes on to paint a daunting outlook for the U.S. if it continues down this perilous road with regard to sustained competitiveness.

As we celebrate National Computer Science Education Week (http://www.csedweek.org/) this month, we must be mindful that recognizing the need to bolster CS in all levels of educational in order for citizens to prosper in the 21st century should be a *daily* mantra...now, more than ever.

ACM Member News

DAWN SONG WINS MACARTHUR 'GENIUS' AWARD

When the John D. and Catherine T. MacArthur Foundation called Dawn Song to tell her she had been awarded one of its prestigious "genius" grants, she almost didn't answer the phone.

"I wasn't really sure whether I was going to pick up the call, because it was from a number I didn't know," says Song, a 35-year-old expert in computer security. When she did answer, she initially thought the foundation was trying to lure her away from her job as an associate professor of computer science at the University of California, Berkeley.

In fact, Song was one of 23 recipients of the $500,000, five-year unrestricted fellowship, designed to let creative people pursue their interests. She says her approach to computer security is fairly broad, studying the underlying structure of computer systems that might lead to vulnerabilities. She's trying to develop a formal description of the Web, the constant evolution of which leads to changing interactions that can open up avenues of attack. Much analysis in computer security is still done manually, and Song wants to reduce the response time to threats by designing tools that automatically detect and defend against attacks.

A growing area of interest for Song is mobile and embedded systems. For instance, medical devices, such as implanted insulin pumps and defibrillators, are connected to wireless networks, providing doctors more information but also creating new vulnerabilities.

The grant, Song says, will allow her to pursue unconventional research. It also gives her the freedom to pursue a major goal of hers. "I believe that life is about creating something truly beautiful," she says. "A piece of good technology, a great scientific finding, I think there is true beauty in that."

—Neil Savage

PHOTOGRAPH COURTESY THE JOHN D. AND CATHERINE T. MACARTHUR FOUNDATION.

Technology | DOI:10.1145/1859204.1859211 Gregory Goth

The Eyes Have It

Eye-tracking control for mobile phones might lead to a new era of context-aware user interfaces.

HUMAN-COMPUTER INTERFACES (HCIs) controlled by the eyes is not a novel concept. Research and development of systems that enable people incapable of operating keyboard- or mouse-based interfaces to use their eyes to control devices goes back at least to the 1970s. However, mass adoption of such interfaces has thus far not been necessary nor pursued by system designers with any particular ardor.

"I'm a little bit of a naysayer in that I think it will be very, very hard to design a general-purpose input that's superior to the traditional keyboard," says Michael Holmes, associate director for insight and research at the Center for Media Design at Ball State University. "When it comes to text, it's hard to beat the speed of a keyboard."

However, one class of user interfaces (UIs) in particular has proven to be problematic in the creation of comfortably sized keyboards—the interfaces on mobile phones, which are becoming increasingly more capable computational platforms as well as communications devices. It might stand to reason that eye-based UIs on mobile phones could provide users with more options for controlling their phones' applications. In fact, Dartmouth College researchers led by computer science professor Andrew Campbell recently demonstrated with their EyePhone project that they could modify existing general-purpose HCI algorithms to operate a Nokia N810 smartphone using only the device's front-facing camera and computational resources. The new algorithms' accuracy rates, however, also demonstrated that the science behind eye-based mobile control needs more refinement before it is ready for mass consumption.

"People tend not to point with their eyes," notes Roel Vertgaal, associate professor of human-computer interaction at Queen's University, who studies eye communication.

PHOTOGRAPH BY JALANI MORGAN

Eyes As an Input Device

Perhaps the primary scientific barrier to reaching a consensus approach to eye-controlled mobile interfaces is the idea that trying to design such an interface flies against the purpose of the eye, according to Roel Vertegaal, asso-

ciate professor of human-computer interaction at Queen's University.

"One of the caveats with eye tracking is the notion you can point at something," Vertegaal says. "We didn't really like that. People tend not to point with their eyes. The eyes are an input device and not an output device for the body."

Vertegaal says this basic incompatibility between the eyes' intended function and the demands of using them as output controllers presents issues including the Midas Touch, postulated by Rob Jacob in a seminal 1991 paper entitled "The Use of Eye Movements in Human-Computer Interaction Techniques: What You Look At is What You Get."

"At first, it is empowering to be able simply to look at what you want and have it happen, rather than having to look at it (as you would anyway) and then point and click it with the mouse or otherwise issue a command," Jacob wrote. "Before long, though, it becomes like the Midas Touch. Everywhere you look, another command is activated; you cannot look anywhere without issuing a command."

Another issue caused by trying to make the eyes perform a task for which they are ill-suited is the lack of a consensus on how best to approach designing an eye-controlled interface. For example, one of the most salient principles of mainstream UI design, Fitts's Law, essentially states that the time to move a hand toward a target is affected by both the distance to a target and the size of the target. However, Fitts's Law has not proven to be a shibboleth among eye-tracking researchers. Many contend the natural accuracy limitations of the eye in pointing to a small object, such as a coordinate on a screen, limit its applicability. A lack of consensus on the scientific foundation of eye control has led to disagreement on how best to approach discrete eye control of a phone. The Dartmouth researchers, for example, used blinks to control the phone in their experiment. However, Munich-based researcher Heiko Drewes found that designing a phone that follows gaze gestures—learned patterns of eye movement that trigger specific applications, rather than blinks—resulted in more accurate responses from the phone.

"I tried the triggering of commands by blinking, but after several hundred blinks my eye got nervous—I had the feeling of tremor in my eyelid," Drewes says. "I did no study on blinking, but in my personal opinion I am very skeptical that blinking is an option for frequent input. Blinking might be suitable for occasional input like accepting an incoming phone call."

However, Drewes believes even gaze gestures will not provide sufficient motivation for mass adoption of eye-controlled mobile phones. "The property of remote control and contact-free input does not bring advantage for a device I hold in my hands," he says. "For these reasons I am skeptical regarding the use of gaze gestures for mobile phones.

"In contrast, I see some chances for controlling a TV set by gaze gestures. In this case the display is in a distance that requires remote control. In addition, the display is big enough that the display corners provide helping points for large-scaled gesture, which are separable from natural eye movements."

Steady Progress

Vertegaal believes the most profound accomplishment of the Dartmouth EyePhone work may be in the researchers' demonstration of a mobile phone's self-contained image and processing power in multiple realistic environments, instead of conducting experiments on a phone tethered to a desktop in a static lab setting. Dartmouth's Campbell concurs to a large degree.

"We did something extremely simple," Campbell says. "We just connected an existing body of work to an extremely popular device, and kind of answered the question of what do we have to do to take these algorithms and make them work in a mobile environment. We also connected the work to an application. Therefore, it was quite a simple demonstration of the idea."

Specifically, Campbell's group used eye-tracking and eye-detection algorithms originally developed for desktop machines and USB cameras. In detecting the eye, the original algorithm produced a number of false positive re-

Obituary

Benoît Mandelbrot, Mathematician, 1924–2010

Benoît Mandelbrot, working largely outside the mainstream of mathematics and computer science, achieved international fame by introducing the term "fractal" to refer to intriguing mathematical shapes that display rough and irregular patterns found in nature. The mathematician died from pancreatic cancer on October 14 at age 85.

Mandelbrot, born in Poland, raised in France, and later a resident of Cambridge, MA, became interested in unusual natural patterns—including coastlines, plants, and blood vessels—as a young man. He continued to focus on geometric complexities throughout his career. In the 1950s, Mandelbrot argued that it was possible to quantify the crookedness of complex objects by assigning a "fractal dimension." In 1982, he published *The Fractal Geometry of Nature*, a book that earned him widespread acclaim.

"His and other mathematicians' work on fractals helped to answer some nagging questions that had bothered mathematicians since the early 20th century," says Joe Warren, professor of computer science at Rice University. In addition, Mandelbrot helped define the fields of computer science, geology, medicine, cosmology, and engineering. In 1979, he studied the Mandelbrot set that is now named after him.

"Movies such as *Avatar* and video games such as *Spore* utilize this method in building realistic creatures and natural environments as well as creating fantastic new creatures and environments," Warren notes. Today, "the study of self-similarity as occurring in fractals also spurred the idea of exploiting self-similarity in other mathematical applications such as signal processing," says Warren.

In 1958, Mandelbrot was hired by IBM to work as a researcher. He served as a visiting professor at Harvard and Massachusetts Institute of Technology universities and accepted a full-time teaching position at Yale University in 1987. Mandelbrot received more than 15 honorary doctorates and served on the board of many scientific journals. He is frequently referred to as the "father of fractal geometry."

—*Samuel Greengard*

sults for eye contours, due to the slight movement of the phone in the user's hand; interestingly, the false positives, all of which were based on coordinates significantly smaller than true eye contours, seemed to closely follow the contours of the user's face. To overcome these false positive results, the Dartmouth researchers created a filtering algorithm that identified the likely size of a legitimate eye contour. The new eye-detection algorithm resulted in accuracy rates of 60% when a user was walking in daylight, to 99% when the phone was steady in daylight. The blink detection algorithm's accuracy rate ranged from 67% to 84% in daylight.

Campbell believes the steady progress in increasing camera resolution and processing capabilities on mobile phones will lead to more accuracy over time. "Things like changes in lighting and movement really destroy some of these existing algorithms," he says. "Solving some of these context problems will allow these ideas to mature, and somebody's going to come along with a really smart idea for it."

However, veterans of eye-tracking research do not foresee a wave of eyes-only mobile device control anytime soon, even with improved algorithms. Instead, the eye-tracking capabilities on mobile devices might become part and parcel of a more context-aware network infrastructure. A phone with eye-gaze context awareness might be able to discern things such as the presence of multiple pairs of eyes watching its screen and provide a way to notify the legitimate user of others reading over his or her shoulder. An e-commerce application might link a user's gaze toward an LED-enabled store window display to a URL of more information about a product or coupon for it on the phone. Campbell says one possible use for such a phone might be in a car, such as a dash-mounted phone that could detect the closing of a drowsy driver's eyes.

Ball State's Holmes says such multimodal concepts are far more realistic than an either/or eye-based input future. "Think about how long people have talked about voice control of computers," he says. "While the technology has gotten better, context is key. In an open office, you don't want to hear everybody talk to their computer. Voice command is useful for things like advancing slide

Eye-based user interfaces on mobile phones could provide users with more options for controlling their phones' applications.

show, but for the most part voice control is a special tool. And while I can see similar situations for eye gaze control, the notion that any one of these alternative input devices will sweep away the rest isn't going to happen. On the other hand, what is exciting is we are moving into a broader range of alternatives, and the quality of those alternatives is improving, so we have more choices." C

Further Reading

Drewes, H.
Eye Gaze Tracking for Human Computer Interaction Dissertation, Ludwig-Maximilians-Universität, Munich, Germany, 2010.

Jacob, R.J.K.
The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Transactions on Information Systems 9*, 2, April 1991.

Majaranta, P. and Räihä, K.-J.
Twenty years of eye typing: systems and design issues. *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, New Orleans, LA, March 25–27, 2002.

Miluzzo, E., Wang, T. and Campbell, A.T.
EyePhone: activating mobile phones with your eyes. *Proceedings of the Second ACM SIGCOMM Workshop on Networking, Systems, and Applications on Mobile Handhelds*, New Delhi, India, August 30, 2010.

Smith, J.D., Vertegaal R., and Sohn, C.
ViewPointer: lightweight calibration-free eye tracking for ubiquitous handsfree deixis. *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology*, Seattle, WA, Oct. 23–26, 2005.

Gregory Goth is an Oakville, CT-based writer who specializes in science and technology.

Career

CS Grads Are Well Paid

As the U.S. economy limps toward a recovery, there's some good news for students launching careers in computer science and engineering: the starting pay for college graduates in those fields is well above that of many other disciplines, according to a new report by *The Wall Street Journal*.

The study, conducted for the *Journal* by PayScale.com, shows that graduates with engineering degrees earned an average starting salary of $56,000 in their first full-time jobs out of college. Computer science graduates were second, earning an average salary of $50,000. By comparison, people who graduated with degrees in communications and English earned $34,000 in their first jobs.

PayScale.com conducted the survey between April and June of this year, which included responses from about 11,000 people who graduated between 1999 and 2010. According to *The Wall Street Journal*, the reported starting pay was adjusted for inflation to make the salaries of graduates from different years comparable.

The survey was conducted as part of *The Wall Street Journal*'s Paths to Professions project, which examined a selection of jobs in careers that were deemed to be satisfying, well paid, and having growth potential.

One of the advantages of having a technical degree is that it can be applied to a variety of fields, such as product marketing and advertising, which can increase the marketability of an engineer or computer science major, says Katy Piotrowski, a career counselor and owner of Career Solutions Group in Fort Collins, CO.

"You can't really wing technical skills, so if there's a technical product that needs to be promoted through a marketing activity, there's more confidence with someone who has a technical pedigree" to be able to understand how the product works, Piotrowski says. While technical skills are not valued in every field, such as social work, they do open a lot of doors, she notes.

—Bob Violino

Science | DOI:10.1145/1859204.1859210 Gary Anthes

Topic Models Vs. Unstructured Data

With topic modeling, scientists can explore and understand huge collections of unlabeled information.

TOPIC MODELING, AN amalgam of ideas drawn from computer science, mathematics, and cognitive science, is evolving rapidly to help users understand and navigate huge stores of unstructured data. Topic models use Bayesian statistics and machine learning to discover the thematic content of unlabeled documents, provide application-specific roadmaps through them, and predict the nature of future documents in a collection. Most often used with text documents, topic models can also be applied to collections of images, music, DNA sequences, and other types of information.

Because topic models can discover the latent, or hidden, structure in documents and establish links between documents, they offer a powerful new way to explore and understand information that might otherwise seem chaotic and unnavigable.

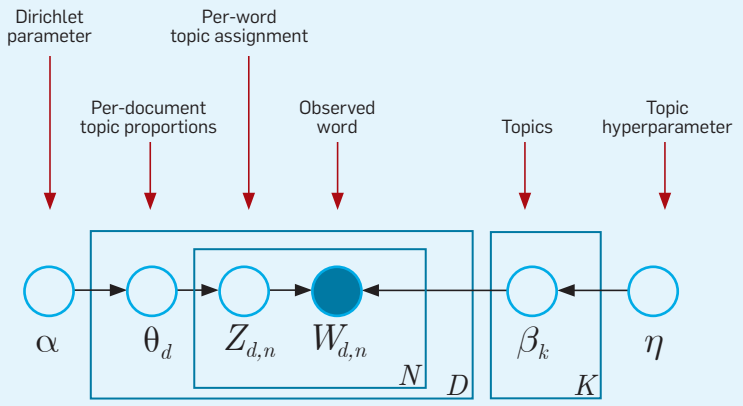
The base on which most probabilistic topic models are built today is latent Dirichlet allocation (LDA). Applied to a collection of text documents, LDA discovers "topics," which are probability distributions over words that co-occur frequently. For example, "software," "algorithm," and "kernel" might be found likely to occur in articles about computer science. LDA also discovers the probability distribution of topics in a document. For example, by examining the word patterns and probabilities, one article might be tagged as 100% about computer science while another might be tagged as 10% computer science and 90% neuroscience.

LDA algorithms are built on assumptions of how a "generative" process might create a collection of documents from these probability distributions. The process does that by first assigning to each document a probability distribution across a small number of topics from among, say, 100 possible topics in the collection. Then, for each of these hypothetical documents, a topic is chosen at random (but weighted by its probability distribution), and a word is generated at random from that topic's probability distribution across the words. This hypothetical process is repeated over and over, each word in a document occurring in proportion to the distribution of topics in the document and the distribution of words in a topic, until all the documents have been generated.

LDA takes that definition of how the documents to be analyzed might have been created, "inverts" the process, and works backward to explain the observed data. This process, called "posterior probabilistic inference," essentially says, "Given these observed data, and given the model for document-creation posited in the generative process, what conditional distribution of words over topics and of topics over documents resulted in the data I see?" It both defines the topics in a collection and explains the proportions of these topics in each document, and in so doing it discovers the underlying semantic structure of the documents.

LDA and its derivatives are examples of unsupervised learning, meaning that the input data is not labeled; the models work with no prior knowledge of the topics in the documents. The models can perform their inference by a number of different algorithms, but they all work by machine learning. They start with random assumptions about the probability distributions, try them out on the data to see how well they fit, then update them and try again.

Latent Dirichlet allocation.



Each piece of the structure is a random variable.

More Modular, More Scalable

LDA is essentially a technical refinement—making it more modular and scalable—of the topic modeling technique called probabilistic latent semantic indexing. Introduced in 1999 by Jan Puzicha and Thomas Hofmann, probabilistic latent semantic indexing was derived from Latent Semantic Indexing, which was developed in the late 1980s by Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas

K. Landauer, and Richard Harshman. Because of its modularity, LDA has become the springboard for a host of refinements and extensions. For example, David Blei, a computer scientist at Princeton University, who co-developed LDA with Andrew Ng and Michael Jordan, has inserted into LDA a stochastic method that relates topics to each other and displays them in potentially infinite tree-like hierarchies, with the various branches located at different levels of abstraction. Unlike LDA, hierarchical LDA does not have to be told the number of topics in advance, and each topic has the potential to break into smaller subtopics without limit.

Blei also codeveloped the correlated topic model (CTM), which can find correlations between topics, which LDA can't do. CTM, for instance, could recognize that the topic "neuroscience" is more closely related to "biology" than it is to "geology."

Another LDA extension, called dynamic topic modeling (DTM), takes into account the evolution of a collection over time. It can recognize that a paper from 1910 on "phrenology" and one from 2010 on "neural networks" might both be classified in the evolving field of "neuroscience," even though they use vastly different vocabularies. "The DTM connects these topics across time so you can see how the lexicon evolved," says DTM pioneer John Lafferty, a professor of computer science, machine learning, and statistics at Carnegie Mellon University.

Researchers at the University of California, Irvine have developed two LDA derivatives that discover the relationships among the entities in news stories—people, organizations, and locations—and the topics found by classical LDA. They applied these "entity-topic models" to a collection of 330,000 *New York Times* articles and found they were especially adept at predicting the occurrence of entities based on the words in a document, after having been trained on test documents. For example, they report that an article about the Sept. 11, 2001 terrorist attacks was likely to include the entities "FBI," "Taliban," and "Washington."

Blei says three developments in the past 10 years have led to rapid advancements in topic modeling: the emergence of LDA as a kind of development platform, advancements in the use of machine learning to perform statistical inference, and "the emergence of very large, unlabeled data sets."

Latent Dirichlet allocation both defines the topics in a collection and explains the proportions of these topics in each document, thereby discovering the underlying semantic structure of the documents.

Despite the models' growing popularity, Blei offers several caveats about their use. He warns against the blind acceptance of results suggested by the models as conclusive. "It is important to be careful," he said. For example, running topic models multiple times, with the algorithms choosing different random initializations of the topics, can lead to very different results. "Also, it can be important to check sensitivity to different choices in the model," says Blei. "There are many dials to tune in topic modeling."

Mark Steyvers, a professor of cognitive sciences at the University of California, Irvine, is exploring how people can analyze documents when they have little idea of what is contained in them. Steyvers and his colleagues have recently used topic models in three real-world situations. In the first, a lawyer had received a large stack of papers related to a lawsuit, and needed a summary picture of their contents. In the second project, funded by a U.S. intelligence agency, the task was to examine huge feeds of email and documents and to provide analysts with lists of their topics. In the third, a government agency wanted to

Milestones

CS Awards

The Glushko-Samuelson Foundation, National Academy of Engineering, and ACM's Special Interest Group on Security, Audit and Control (SIGSAC) recently recognized leading computer scientists for their research and leadership.

DAVID E. RUMELHART PRIZE

Judea Pearl, director of the Cognitive Systems Laboratory in the department of computer science at the University of California Los Angeles, is the recipient of the eleventh David E. Rumelhart Prize. The prize is awarded annually by the Glushko-Samuelson Foundation to an individual or collaborative team making a significant contemporary contribution to the theoretical foundations of human cognition. Pearl has developed Bayesian networks that can be used to represent and draw inferences from probabilistic knowledge in a highly transparent and computationally tractable fashion.

ARTHUR M. BUECHE AWARD

Anita Jones, a university professor emerita in the computer science department at the University of Virginia, was awarded the Arthur M. Bueche Award from the National Academy of Engineering for "leadership in the development of U.S. science and technology policy and the development of technologies for national security, including technical contributions to high-performance computing and cybersecurity."

SIGSAC AWARDS

SIGSAC presented its top honors to Jan Camenisch of IBM Research-Zurich and Bhavani Thuraisingham of the University of Texas at Dallas for their contributions to the computer and communications security community. Camenisch received the SIGSAC Outstanding Innovation Award for his theoretical work on privacy-enhancing cryptographic protocols and his leadership in their practical realization. Thuraisingham received the SIGSAC Outstanding Contribution Award for her seminal research contributions and leadership in data and applications security over the past 25 years.

—Jack Rosenberger

understand the topics and inter-topic relationships among the hundreds of thousands of grants awarded by it and sister agencies.

The ultimate application may be to help understand how the human mind works. Steyvers is experimenting with topic modeling to shed light on how humans retrieve words from memory, based on associations with other words. He runs the models on educational documents to produce crude approximations of the topics learned by students, then compares the accuracy of recall, based on word associations, of the students and models. Sometimes the models make mistakes in their word and topic associations, which are shedding light on the memory mistakes of humans. What's needed, Steyvers says, is nothing less than "a model of the human mind."

Meanwhile, computer scientists are looking for ways to make algorithms more efficient and to structure problems for parallel processing, so that huge problems, such as topic modeling the entire World Wide Web, can be run on large clusters of computers.

Fernando Pereira, a research director at Google, says a number of experimental systems of probabilistic topic modeling are being investigated at the company. The systems could provide better Google search results by grouping similar terms based on context. A topic model might discover, for instance, that a search for the word "parts," used in an automobile context, should include "accessories" when it is also used in an automobile context. (The two words are seen as synonyms if both are used in the same context; in this case, automobiles.) Google does some of that now on a limited basis using heuristic models, but they tend to require a great deal of testing and tuning, Pereira says.

One of the advantages of the LDA framework is the ease with which one can define new models.

"I can't point to a major success yet with the LDA-type models, partly because the inference is very expensive," Pereira says. "While they are intriguing, we haven't yet gotten to the point that we can say, 'Yes, this is a practical tool.' "

But, says Tom Griffiths, director of the Computational Cognitive Science Lab at University of California, Berkeley, "We are seeing a massive growth in people applying these models to new problems. One of the advantages of this [LDA] framework is it's pretty easy to define new models."

Further Reading

Blei, D. and Lafferty, J.
Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, June 25–29, 2006.

Blei, D. and Lafferty, J.
Topic models, *Text Mining: Classification, Clustering, and Applications*, (Srivastava, A. and Sahami, M., Eds), Taylor & Francis, London, England, 2009.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D.
Reading tea leaves: How humans interpret topic models. Twenty-Third Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, Dec. 7–12, 2009.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R.
Indexing by latent semantic analysis, *Journal of the American Society for Information Science 41*, 6, 1990.

Newman, D., Chemudugunta, C., Smyth, P., and Steyvers, M.
Statistical entity-topic models. The Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, PA, August 23–26, 2006.

Gary Anthes is a technology writer and editor based in Arlington, VA.

History

Building Babbage's Analytical Engine

A British programmer and author who wants to build a computer based on 19th century designs has about 3,000 people pledging donations to his project. John Graham-Cumming, author of *The Geek Atlas*, hopes to build an Analytical Engine invented by English mathematician Charles Babbage and first proposed in 1837. Babbage is often called "the father of computing."

The Analytical Engine, originally meant to be constructed of iron and brass and operated with steam power, will have the equivalent of 1.7 kilobyte of memory and be capable of four arithmetic operations: left and right shift, and comparison/jump operations. It will be very slow, with a single addition taking about 13,000 times as long as on a Z80, an 8-bit microprocessor from the mid-1970s.

"I think it would be an inspirational machine to see," Graham-Cumming said via email. "People could literally see how a computer operates since the architecture of the machine is close to the architecture of modern computers.

"It will answer the question: Could Victorians have had the computer? Others will have to answer the question: And what difference would that have made?" he says.

One challenge will be deciding what version of the machine to build as Babbage was continually refining his designs. Graham-Cumming is focusing his efforts on a version called Plan 28, but says more research is needed. Eventually, he plans to create a 3D virtual model of the Analytical Engine, work out all the bugs, and build it. As part of the project, Babbage's papers will be digitized and made available online.

The project could take up to five years and cost £1 million (about $1.6 million U.S.). Graham-Cumming has started a Web site, http://www.plan28.org, to solicit donations, and plans to start work once he has 10,000 pledges. Another Babbage machine, a calculator called the Difference Engine No. 2, was built by the London Science Museum in 1991.

In 2009, Graham-Cumming led an online petition demanding an apology from the British government for its treatment of World War II mathematician and code-breaker Alan Turing, who was prosecuted for being a homosexual. The petition, which attracted international media attention, was successful.

—Neil Savage

Education | DOI:10.1145/1859204.1859213 Marina Krakovsky

CSEdWeek Expands Its Reach

The second Computer Science Education Week is showing students, parents, and educators why computer science is important.

TO GRADUATE FROM the Advanced Math and Science Academy, a charter school for grades 6–12 in Marlborough, MA, every high-school student must take at least three years of computer science. However, the public school's inclusion of computer science alongside math, English, and other core subjects is a remarkable exception in the U.S., where only one in five states counts computer science classes toward any kind of graduation requirement. In fact, even as most occupations are increasingly dependent on computing, the number of computer science courses in the U.S. has decreased in the past five years, according to *Running on Empty: The Failure to Teach K–12 Computer Science in the Digital Age*, a new report from ACM and the Computer Science Teachers Association. The report (http://www.acm.org/runningonempty/) also notes that when schools offer a computer science course, it is usually an elective; moreover, much of what passes for high-school computer science instruction is actually about information technology (IT) literacy rather than algorithm design, programming, or computational thinking.

"Computing underpins everything, yet two-thirds of states have few learning standards in the fundamentals of computer science," says ACM CEO John White. In addition to its efforts at state and federal levels, ACM is leading the second Computer Science Education Week (CSEdWeek), which is aimed directly at students, teachers, parents, and counselors this year. Held Dec. 5–11 and officially recognized by a Congressional resolution, CSEdWeek is spreading the message that computer science is a crucial part of a 21st century education—regardless of your future career.

"There is a growing misconception that if you can send an email or find something on the Web that you somehow know a lot about computing," says White. "No, you're increasingly IT literate, but you know nothing about the fundamentals of how this stuff works."

Bolstering the teaching of mathematics is not enough, says White, explaining that computer science is its own discipline, distinct from every dimension of math taught in high school. "CS has a lot of math in it, but it's a lot more than mathematics," he says. "It's one thing to develop a static mathematical solution, but it's another to build a dynamic computation to achieve whatever your goal is."

Supported by a group of corporate and nonprofit partners, this year's CSEdWeek budget is twice that of last year, White says, and some of it has been spent on a richer Web site, http://www.csedweek.org, which conveys the importance of computer science, offers curriculum materials, and suggests ways to participate. Local events throughout the U.S., Canada, and other countries range from CS Unplugged activities and computing diaries to field trips and school visits from ACM chapters.

CSEdWeek's organizers' long-term goal is to make the event as prominent and influential as its older cousin, Engineers Week. "Engineering, like computing, is a discipline that's not part of the core in K–12 education," says Debra Richardson, professor of informatics at the University of California, Irvine and chair of this year's CSEdWeek. "But nobody thinks there are no jobs in engineering, whereas there's still this sense on the part of parents and guidance counselors that there are no jobs in computing."

Although it's true that many computing-related jobs have gone overseas, Richardson says these are mainly jobs in customer service and quality assurance. "Computer science trains you for developing the next intellectual property for a company, and they're not offshoring that," says Richardson, citing the U.S. Department of Labor's Bureau of Labor Statistics projection of 800,000 to 1.5 million U.S. computing jobs between now and 2018.

Moreover, Richardson and White stress that computer science is invaluable for innovation in fields as disparate as biology, engineering, and health care. "Studying computer science means you'll be able to use computation in solving problems," says White, "and that's a huge amplifier in our ability to advance almost any aspect of society." C

Marina Krakovsky is a San Francisco area-based journalist and co-author of *Secrets of the Moneylab: How Behavioral Economics Can Improve Your Business.*

Society | DOI:10.1145/1859204.1859212 Samuel Greengard

The New Face of War

With the introduction of the sophisticated Stuxnet worm, the stakes of cyberwarfare have increased immeasurably.

Ever since European cybersecurity officials discovered the Stuxnet worm last June, it has been characterized as a "paradigm shift" in critical infrastructure threats. European Network and Information Security Agency Executive Director Udo Helmbrecht characterized Stuxnet, which is unprecedented in its capabilities and sophistication, as "a new class and dimension of malware."

Stuxnet, which contains four zero-day Windows vulnerabilities as well as two stolen digital certificates for authentication is the first discovered worm that secretly monitors and reprograms industrial control systems. Stuxnet exploits weaknesses in Windows operating systems and takes command of a Siemens component that controls critical industrial operations, such as those of oil pipelines, electrical power grids, and nuclear energy plants.

Whether Stuxnet is a new weapon of modern espionage or cyberwarfare is unclear. However, many security experts believe the sophisticated malware was developed by a well-funded private entity or a national government agency to attack Iran's industrial infrastructure, including the Bushehr nuclear power plant. Iranian officials report that Stuxnet has infected 30,000 machines involved in running its industrial control systems, and the Bushehr facility reportedly didn't work correctly for several months. "An electronic war has been launched against Iran," according to Mahmoud Liaii, director of Iran's Information and Technology Council of the Industries and Mines Ministry.

To be certain, the digital age is ushering in entirely new ways to fight wars. "Cyber tools can be used as an instrument for government security as well as for military and intelligence purposes," states Herbert Lin, chief scientist for the Computer Science and Telecommunications Board at the U.S. National Research Council. Adds Rain Ottis, a staff scientist at the Cooperative Cyber Defence Centre of Excellence in Tallinn, Estonia, "Cyber warfare is almost certain to emerge the next time two technologically advanced states fight a major shooting war."

Cyberwarfare or industrial espionage? The Stuxnet worm has infected 30,000 machines in Iran, including an unknown number at the Bushehr nuclear power plant.

It's not an abstract concept. Although government Web sites and computer systems are likely targets (many government systems are 10 to 20 years old and can't support modern security standards), civilian targets like power grids, telecommunications networks, flight control systems, and financial networks are also at risk. "Cyberwarfare has the potential to cause significant strategic damage," observes Sami Saydjari, CEO of Cyber Defense Agency, a cybersecurity consulting firm headquartered in Wisconsin Rapids, WI.

The Changing Nature of War

The evolution of weaponry has always centered on gaining superiority over an enemy. However, in the digital age, the nature of war is changing radically. It's one thing to detect an invading army and its well-marked planes, tanks, and troops. It's not so simple to identify bits and bytes of data, ascertain exactly where they're coming from, and understand the sender's intentions.

In fact, hackers and others attack government and corporate systems for a broad array of reasons that have nothing to do with politics or ideology. As a result, defining cyberwarfare is a challenge and the hype often exceeds reality. "Some very visible and vocal people have seemingly equated any malicious activity on the Internet to cyberwarfare," Ottis observes. "Most, if not all, cyberattacks should be classified as criminal, espionage related, or hactivist, and not warfare."

Moreover, there is no clear international law covering cyberattacks although experts say the mere act of one nation or state invading another's computers could be construed as an act of war. Common attack methods include vandalism, spreading propaganda, gathering classified data, using distributed denial-of-service (DDoS) attacks to shut down systems, destroying equip-

PHOTOGRAPH BY ISNA. MEHDI GHASEMI/AP PHOTO

ment, attacking critical infrastructure, and planting malicious software.

Over the last few years, untold incidents may have fallen into these categories, but proving the legitimacy of attacks is next to impossible. That's because hackers hijack computers all over the world and use them as part of a botnet to launch attacks. Tracing back the Internet protocol address doesn't necessarily provide insight into who actually is launching the attack. In May 2007, for example, Estonia's government Web sites came under attack—presumably from Russia—after the government moved a Soviet war memorial. The wave after wave of DDoS attacks came from IP addresses all over the world, though many of them originated from Russian-hosted servers.

In October of the same year, Israel mounted a sneak air attack against Syria and destroyed a fledgling nuclear research center deep inside the country. Some analysts believe that Israel avoided detection by hacking radar and other defense systems in Syria and perhaps additional countries. Afterward, *Aviation Week* magazine reported: "The process involves locating enemy emitters with great precision and then directing data streams into them that can include false targets and misleading message algorithms."

When Russian troops invaded the Republic of Georgia in August 2008 the news media diligently reported the event and analysts pondered the repercussions. But what wasn't apparent to many—at least immediately—was that the battle wasn't being fought only with troops and tanks. Several servers and Web sites operated by the Georgian government, including the nation's primary government site, were rendered useless through a steady barrage of DDoS attacks.

Almost immediately, the Georgian Ministry of Foreign Affairs issued a statement via a replacement site built on a blog-hosting service. It read: *A cyber warfare campaign by Russia is seriously disrupting many Georgian websites*. Meanwhile, Barack Obama, then a U.S. presidential candidate, issued a demand that Russia stop interfering with the Web sites. Analysts and security experts noted that the attacks—mostly originating in Russia and Turkey—were linked to the Russian Business Network, a group with close ties to Russian gangs and the government.

"Cyberwarfare is almost certain to emerge the next time two technologically advanced states fight a major shooting war," says Rain Ottis.

Meanwhile, the People's Republic of China and the U.S. have reportedly launched cyberattacks against each other dating back to the 1990s, though China has taken a lead in developing cyberwarfare systems, Saydjari says. For one thing, the government has adopted a more secure operating system named Kylin, which provides hardened protection that is not available with Windows, Unix, and Linux. China has also funneled capital and expertise into developing cyberwar capabilities, including enlisting patriotic hacker gangs. "They have acted slowly, patiently, and strategically," says Saydjari.

The New Battlefield

Although government-sponsored cyberattacks have so far occurred on a limited basis, the probability of a major cyberwar erupting over the next decade seems inevitable. In all likelihood, experts say, a cyberassault would accompany more traditional forms of warfare, but it could also serve as a way to wreak economic harm or destabilize a nation state without a conventional battle. As Ottis puts it, "In the end, the aim of war is usually not to kill your enemy but to impose your will on them."

Scott Borg, director and chief economist of the nonprofit U.S. Cyber Consequences Unit, located in Norwich, VT, has stated publicly that cyberattacks can cause "horrendous damage." Even a short-lived Internet failure could have severe repercussions. The cost of a flight control system crashing or an electrical power grid fading to black

Cybersecurity

Isolate Infected PCs?

Computers infected with malware should be disconnected from the Internet to prevent them from harming other members of the online community, Scott Charney, corporate vice president of Trustworthy Computing at Microsoft, said during his speech at ISSE 2010. The proposed measure would not only prevent the spread of malware, but also pose substantial difficulties for botnets, Charney noted.

Charney's speech, along with a simultaneously published paper, "Collective Defense: Applying Public Health Models to the Internet," urged the IT security community to rethink its approach to cybersecurity and adopt quarantine measures similar to those adopted by the public-health professionals.

"For a society to be healthy, its members must be aware of basic health risks and be educated on how to avoid them," Charney wrote in the paper. "In the physical world, there are also international, national, and local health systems that identify, track, and control the spread of disease including, where necessary, quarantining people to avoid the infection of others."

Meanwhile, the U.S. government is studying a number of voluntary ways to help the public and small businesses better protect themselves online. The possibilities include provisions in an Australian program that enable customers to receive alerts from their Internet service providers if their computer is hijacked via a botnet. U.S. officials are not advocating an option in the program that permits ISPs to block or limit Internet access by customers who fail to fix their infected computers. However, Harris Corporation's Dale Meyerrose, vice president of Cyber and Information Assurance, warns that voluntary programs will be insufficient. "We need to have things that have more teeth in them, like standards," Meyerrose says.

—Phil Scott

could extend into the hundreds of billions of dollars and lead to a cascade of economic problems.

Yet the ensuing fallout could cause an additional array of headaches. Because China manufactures many components, obtaining spare parts during a conflict could prove difficult, if not impossible. In addition, some analysts worry that electricity generators and other components imported from China and other countries could contain hidden software that allows hackers to access systems through a back door or execute a malicious software program on command.

Already, many nations have developed sophisticated hacking and intrusion capabilities, including planting Trojan horses, rootkits, and other nefarious tools on targeted systems. Many of these applications stealthily reside on computers until someone decides to flip a switch and activate them. In fact, much of the preparation goes on behind the scenes. "During 'peacetime' many nations actively prepare for offensive cyberoperations and some nations probably test their capabilities with nonattributable events," Ottis points out.

Of course, the idea of cyberwarfare hasn't been lost on terrorist organizations either. "While nation-states are likely to be quite choosey about how and when they use cyberwarfare tools, terrorists are likely to view things in a less methodical and calculating way," Saydjari explains. "Their goal can be as simple as destabilizing systems and creating chaos." Worse, it is nearly impossible to identify terrorists inflicting a cyberattack and strike back at a tangible target. Economic sanctions and diplomacy aren't viable either.

China has funneled capital and expertise into developing its cyberwar capabilities, including enlisting civilians and gangs of hackers.

What makes the situation all the more dangerous, Saydjari notes, is that a relatively large number of cybermercenaries exist and many openly advertise their skills. In some instances, they might be hired to handle a project that appears less harmful than it actually is or they might not be concerned about the repercussions. In addition, these individuals often act in a rogue manner and they can easily venture beyond a government's desired actions.

There is some pushback. For now, some businesses and governments are turning to ethical hackers to discover holes and vulnerabilities and report them to authorities. In addition, the U.S. government recently announced a $40-billion national cybersecurity plan to combat cyberattacks from foreign and domestic hackers. However, in May 2010, James Miller, principal deputy under secretary of defense for policy for the U.S. Department of Defense, noted the nation is losing enough data from cyberattacks to fill the Library of Congress many times over.

Make no mistake, the risk of cyberwarfare is growing and many, including Saydjari, warn that political leaders aren't entirely tuned into the severity of the threat. Murky definitions, old and insecure computer systems, difficult-to-detect actions, and vague rules of engagement aren't making things any easier. "Cyberwarfare must be taken seriously," Lin says. "The question isn't, Will it happen? It's *how* and *when* will it happen and what effect will it have on society." C

Further Reading

Bruno, G.
***The Evolution of Cyber Warfare*, Council on Foreign Relations, Feb. 2008.**

Clarke, R.A. and Knake, R.
***Cyber War: The Next Threat to National Security and What to Do About It,* Ecco, New York, NY, 2010.**

Drogin, B.
"In a doomsday cyber attack scenario, answers are unsettling," *Los Angeles Times*, February 17, 2010.

Mueller, R.S. III.
"The State of Cyberterrorism and Cyberattacks," speech, RSA Cyber Security Conference, San Francisco, CA, March 4, 2010.

National Research Council
***Technology, Policy, Law, and Ethics Regarding U.S. Acquisition and Use of Cyberattack Capabilities*, 2009.**

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

Security

India Plans Its Own OS

India recently announced plans to develop its own operating system for government and high-level corporate computers to make them more secure.

Concerns about the vulnerabilities of Western-developed OSs and software and of cyberattacks from computers based in China raised the issue to one requiring a significant response from the Indian government as it moves forward technologically. Last year China developed its own OS, Kylin, for government computers.

India's Defense Research and Development Organization (DRDO) will partner with several other groups, including the Indian Institute of Science and the Indian Institute of Technology Madras, to build the new OS.

"With a homegrown system, the source code will be with us and it helps in securing our systems," according to Vijay Kumar Saraswat, scientific adviser to the defense minister and DRDO director-general. Development and maintenance of the OS source code is considered vital to protecting India's computer networks, Saraswat said.

However, some critics question the usefulness of an Indian OS and the government's motives. It appears Indian officials plan to keep the source code secure by not letting the OS be widely used beyond the nation's borders, but Bruce Schneier, chief security technology officer at BT, suspects this security measure will eventually fail, virtually making the OS irrelevant.

Indian officials have not made cost or time estimates for the project, leading some to suggest it is a public-relations ploy.

—Graeme Stemp-Morlock

Forum | DOI:10.1145/1859204.1859214 David Lindley

A Matter of Privacy

Do consumers have enough control over their personal information or is more government regulation needed?

Google presents you with ads related to your search. Amazon asks if you're interested in a new camera. LinkedIn comes up with names and faces of people it thinks you might know. You may find this very useful—or annoying, or even disturbing. The more we live on the Internet, the more it knows about us, but how much do we want it to know? And who gets to decide? Those were central but largely unanswered questions that panelists wrestled with at a forum on "The Future of Privacy Online," organized by the Information Technology & Innovation Foundation (ITIF) and the Technology Policy Institute (TPI) in Washington, D.C., on September 27.

Whatever concerns people may have over the privacy of data have generally not been enough to deter them from posting all manner of personal information on social networks. Nor do consumers turn away from shopping Web sites that record user preferences in order to generate targeted advertising. In fact, as ITIF Senior Analyst Daniel Castro explained, advertisers are willing to pay twice as much or more for targeted advertising because they get better consumer responses. Castro and other panelists also discussed a recent paper, "Privacy Regulation and Online Advertising," by Avi Goldfarb of the University of Toronto and Catherine Tucker of the Massachusetts Institute of Technology, which showed that a European Union directive limiting the ability of Web site operators to gather and use personal information reduced the effectiveness of targeted advertising, measured by the change in consumers' stated intention to purchase, by as much as 65%.

Data regulation to protect privacy translates into costs to businesses, said TPI President Thomas Lenard, while it remains unclear if there is any compensatory benefit to the consumer. Moreover, excessive regulation can stifle innovation. Panelists expressed a preference for either voluntary codes of practice or regulation along the lines of the Fair Credit Reporting Act, which strengthened the U.S. credit industry less by limiting the kinds of data that can be collected than by setting clear rules for its use. As National Telecommunications and Information Administration Associate Administrator Daniel Weitzner put it, transparent regulations tell businesses what they can and cannot do, while protecting the consumer through provisions that "only kick in when something bad happens."

Facebook Director of Public Policy Tim Sparapani, who spoke at the forum on "The Future of Privacy Online" in Washington, D.C.

Mark Eichorn, assistant director at the Federal Trade Commission, argued that users should have the means to know that personal information is being collected about them and be given choices. That point was echoed by Tim Sparapani, director of public policy at Facebook, who said that one-size-fits-all regulations would hamper businesses, and that privacy controls should be consumer driven. In particular, Sparapani said, increasingly mobility of Internet access—meaning that a person might access financial services through a friend's cell phone—required flexibility even in an individual's choice of privacy rules. Eichorn cautioned, however, that the proliferation of systems for navigating and selecting privacy options forces users to do a lot of work, and may cause some to just give up and go with defaults.

In the end, the overriding sentiment at the forum was that privacy concerns are best dealt with by giving consumers the power of choice and letting the market decide. But "this marketplace is in its infancy," as Sparapani noted, and it's not at all clear that consumers are in a position to understand the consequences of their choices.

David Lindley is a science writer and author based in Alexandria, VA.

PHOTOGRAPH BY ANDREA BRUCE/WASHINGTON POST/GETTY IMAGES

DOI:10.1145/1859204.1859215

Seymour Goodman and Andrew Harris

Emerging Markets
The Coming African Tsunami of Information Insecurity

As the affordability and use of mobile phones in Africa increase, so too will security vulnerabilities.

Over the course of the past decade far too many African nations have continued to struggle, plagued by famine, disease, and conflict. However, there has been one consistently positive African story: the astounding diffusion of mobile phones across the continent. With current growth rates more than twice as fast as the rest of the world, Africans have embraced the cellular device to a degree once thought unimaginable. The world's second most populous continent now faces the very real possibility of capturing the potential of mobile telephony and launching a new era of economic and social development.

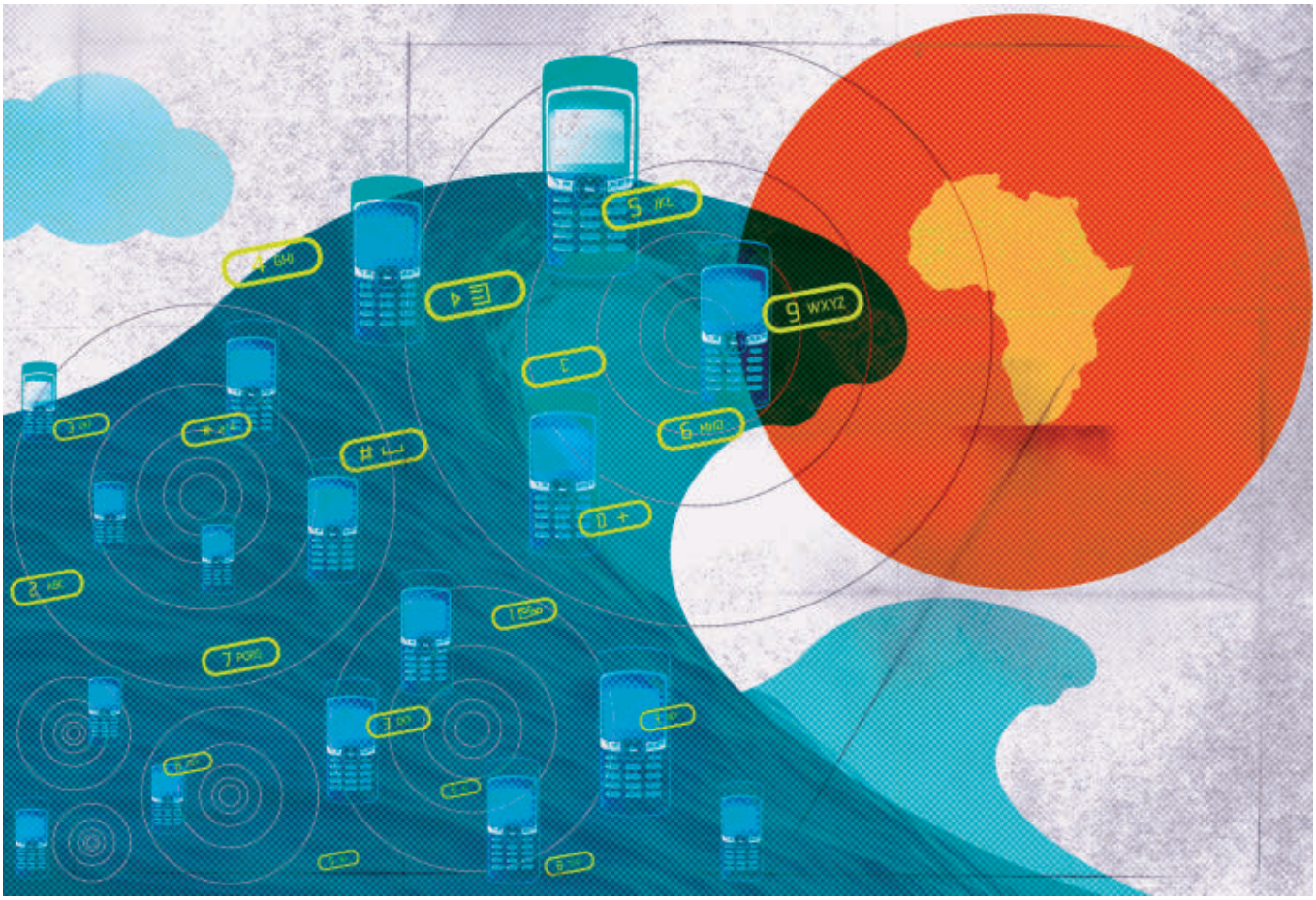
This is not a new story, but it is a story only partly considered. For all of the value these devices will deliver, a darker side of the wireless revolution remains rarely discussed. As the use and utility of mobile phones in Africa continues to rise, so too will security vulnerabilities. Unless properly addressed, security vulnerabilities endemic to the use of the information and communications technologies (ICTs) will be magnified by a number of factors unique to Africa, possibly leading to a tsunami of information insecurity across the continent.

Rapid Adoption Rate

It is important to note that the rapid growth of mobile phones stands in contrast to the much slower adoption of other ICTs. In fact, in 2009 Africa boasted 295 million mobile phone subscriptions for a penetration rate of 37.5 per 100 inhabitants, compared to just 8.8 per 100 Internet subscribers and less than two per 100 landline telephone users.[5] Africa's annual growth rate of 47% between 2003–2008 far exceeds the worldwide rate of 21.5%, demonstrating the extraordinary continental adoption of mobile phones.[4] A number of supply and demand factors have propelled this growth, such as the availability of prepaid airtime options, the existence of a competitive market in nearly every African nation, and the limited infrastructure requirements of establishing a cellular network. In contrast, market structure, user costs, and lack of infrastructure will continue to restrain the use of Internet and landlines. Meanwhile, the affordability and growing utility of mobile phones will make the device an increasingly essential tool for the average African.

Beyond assisting in the modernization of African economies, mobile phones have provided welcome security uses as well.

Africans have proven innovative with mobile phones by creating unique banking and security applications as well as adopting low-cost pricing schemes that allow greater access to poor populations. Perhaps the most compelling and fastest-



growing African mobile application is mobile banking. Previously a continent of largely unbanked populations, service providers are rapidly adding low-cost mobile banking solutions for their customers. One of the leaders in this important area is Kenyan mobile service provider, Safaricom. Its M-PESA mobile banking service obviates the need for a physical branch, allowing users to conveniently save, transfer, and spend large and small amounts of cash no matter their physical location or economic status. Rather than simply offering a mobile portal for existing bank customers, as is common in Western nations, M-PESA has created new customers from a population previously denied banking services. The continued growth of mobile banking has the potential to rapidly modernize commerce and personal finance in Africa.

Beyond assisting in the modernization of African economies, mobile phones have provided welcome security uses as well. Following post-election violence in 2008, Kenyans were able to send SMS alerts regarding outbreaks of violence, thus allowing officials and activists to accurately track and respond to threatening situations (see http://www.ushahidi.com/). In Tanzania, police provided free mobile phones to albino citizens in response to targeted murders against that population. The phones were programmed with a special number to directly contact the police allowing users to report threats.[1]

In a continent where many households have little discretionary spending, innovation in the payment of mobile services has been essential to rapid diffusion. Throughout the developing world, prepaid phone service has propelled mobile growth, while in Africa a number of steps have further assisted this growth. Rather than billing by the minute, many providers instead bill by the second. Provider MTN employs a system of dynamic tariff charges in which costs are adjusted each hour depending on typical usage. With the ability to check the cost of calls at any particular time, customers can choose the least expensive times to use their phones.[7]

These examples of low-cost access, security, and financial applications are part of a greater African trend that will only continue as the device becomes more powerful. There is then little doubt that mobile phones will become the primary personal and professional ICT vehicle for a large majority of Africans. This stands in stark contrast to the past when only a small minority of Africans had any regular personal ICT access at all.

Security Vulnerabilities

Certainly the potential gains for Africa are great, but a growing dependence on mobile phones elicits cause for concern. Worldwide trends indicate significant security problems for mobile telephony in the coming years. Mobile devices and networks will become increasingly vulnerable in every way that networked desktops and laptops currently are. Particularly as mobile phones become more vital to personal commerce and finance, the devices

ILLUSTRATION BY STUART BRADFORD

will become more desirable targets for criminals. The botnets and malware that currently plague networked computers will soon become commonplace on our mobile devices, opening the door to data theft and denial-of-service attacks.[3] Limitations of both battery and processing power make mobile phones less defensively capable than desktops and laptops, thus increasing the vulnerability of phones. Beyond cyber attacks, the size and manner of use of mobile phones makes them particularly susceptible to loss and theft. A lost or stolen phone can be mined for personal data or used in a number of malicious ways. This global trend of insecurity applies to African phones as well. Initially, African users will be protected by their relative disadvantages, such as less capable phones and use limitations due to electric power deficiencies. However, as both use and capability of African mobile phones increases, so will criminal activity. While all users of mobile phones should expect a surge of telephonic cyber crime, the state of African information security is likely to increase the vulnerability of African users.

With mobile phones vulnerable to cyber and physical attacks, African users can expect to experience the same set of headaches that are becoming more common in Western nations. Identity theft is typically not a great threat in a developing rural community. However, once individuals begin sharing personal information with service providers, that unique identity becomes vulnerable to theft and misuse. Further, for those participating in mobile banking, a perpetrator who has gained cyber or physical control of the mobile device can obtain complete access to financial records and the ability to conduct transactions. Such a violation could potentially wipe out the savings of a family, leading to years of financial hardship. Mobile phones also provide the potential for snooping, in which perpetrators can listen to conversations and track locations. These are new problems for Africans; never before could an individual's privacy, identity, or savings be compromised because of one device's vulnerabilities.

Because the potential gains are so great for Africa, it is vital that malevolent forces do not spoil this moment of opportunity.

Insecurity Factors

Three factors in particular portend the tide of insecurity. First, the vast majority of African nations suffer from a deficiency of appropriate laws and organizations needed to confront cyber crime. It is only recently the case that both Internet and mobile telephony have come to every African country. This fact coupled with inadequate resources has left most African nations without proper institutions needed to secure the cyber realm. While Tunisia stands out in its institutional readiness with an established national Computer Emergency Response Team (CERT) that has actively reached out to both businesses and the general population, only a few of the remainder of African states have information security teams, and those remain in embryonic stages. This has lead to a high rate of computers infected with malicious software—perhaps as high as 80%.[2] The inability to stop criminal activity now amidst an environment of few sophisticated Internet users speaks to the great problem ahead when many mobile phone users begin to apply their phones in increasingly sophisticated and sensitive ways.

A second issue revolves around African notions of privacy. While the African states are certainly not monolithic in their thinking, there does exist a pervasive ethos of communitarianism that deemphasizes the individual right of privacy. In Europe and North America, the right of personal privacy is far more prevalent, leading to significant advocacy for the protection of personal information in the digital realm. Although there is arguably a great deal more that could and should

be done, this advocacy has forced corporate gatherers of personal information to be mindful of protecting data from misuse or theft, leaving users more protected. Without such a strong notion of privacy rights in Africa, this advocacy is disturbingly absent in the nascent field of African information security. For instance, in 2004 the South African Post Office decided to sell the personal information of citizens in its database. Without any legal mechanism to protect personal information, individuals had no means of protecting their own privacy.[6] Few protections of personal information mean that sensitive data can all too easily fall into the wrong hands.

Finally, too many African governments demonstrate a willingness to operate outside the rule of law and with little accountability. In such an environment, mobile phones become an unprecedented tool to track a citizen's activities. An unscrupulous government could easily use the cellular network to track an individual's movement, listen to conversations, and access financial records. While such behavior is not absent in Europe and North America, it is generally limited due to robust legal systems and privacy watchdogs. Where such systems are absent, as in many African states, government snooping can have a chilling effect on a population and negate the many gains provided by mobile service.

Because the potential gains are so great for Africa, it is vital that malevolent forces do not spoil this moment of opportunity. Further, we must not forget that in such an interconnected world, a problem for Africa is very much a problem for everyone else. Once a continent with very limited broadband connectivity, undersea fiber cables now span the length of each of Africa's coasts and the development of several new cables is under way. While desperately needed, this additional bandwidth can serve as a conduit to import and export mobile viruses and other forms of malware from and to the rest of the world. It is, therefore, imperative that this potential information security nightmare be addressed. Like so many security problems in the cyber world, however, the solutions are not evidently at hand.

The compelling rise of mobile telephony across Africa is not a passing phenomenon.

Of particular concern is the current lack of information security professionals in Africa. What is a significant problem in developing countries is compounded in Africa where few countries have the resources to educate and train the work force needed to protect the cellular networks. Along the same lines, governmental capacity to develop the necessary institutions is largely missing. Western assistance, when offered, is rarely given in a sustainable manner that would allow for true security work to persist. Finally, political sensitivities tend to limit the amount of assistance home nations are willing to accept relating to security issues. All of this leaves Africa with an absence of internal capacity coupled with weak outside assistance. It is therefore difficult to imagine positive scenarios in which the onslaught can be avoided.

Conclusion

In considering possible solutions, it is clear that device manufacturers and service providers must contribute. Too often private interests place security low on the list of priorities, especially when not encouraged by government entities. Yet for African networks to be safe, it will be essential for manufacturers and providers to offer adequate security. Steps such as encrypting all sensitive data passed through the network and ensuring the privacy of personal information offer adequate user protection. African governments must also find ways to train information security professionals. While increasing this work force will certainly carry a high cost to governments that often have little money to spend, the costs otherwise will be far greater, although distributed more broadly among the population.

Beyond Africa, mobile phone security must be raised in the consciousness of Western professionals and the international bodies studying and working in the field of information security. By raising awareness across the globe, we have a far greater chance of motivating sustainable international assistance.

Finally, an African public awareness campaign is vital. As individuals begin to use more powerful devices in more powerful ways, it is essential that these users understand the potentially detrimental effects that lurk unseen.

The compelling rise of mobile telephony across Africa is not a passing phenomenon. While restraints for future growth do exist, the power, accessibility, and affordability of the devices make them an irresistible force in the coming decade. As stated, not only will subscriber numbers increase, but so too will the capability and utility of the devices. Very soon a majority of Africans will be using mobile phones for banking, accessing the Internet, facilitating commerce, and general communication. It is possible the prospect of a tsunami of information insecurity might recede, but this will only occur with early, concerted, and cooperative engagement on behalf of national governments, international donors, device manufacturers, and service providers. C

References

1. A horrendous trade. *The Economist*. (Jan. 17, 2009), p. 50.
2. Gady, F. Africa's cyber WMD. *Foreign Policy* (Mar. 24, 2010).
3. Georgia Tech Information Security Center. *Emerging Cyber Threats Report for 2009*. (Oct. 15, 2008); http://www.gtisc.gatech.edu/pdf/CyberThreatsReport2009.pdf.
4. ITU. *Information Society Statistical Profiles 2009: Africa*; http://www.itu.int/dms_pub/itu-d/opb/ind/D-IND-RPM.AF-2009-PDF-E.pdf.
5. ITU. *Key Global Telecom Indicators for the World Telecommunication Service Sector*; http://www.itu.int/ITU-D/ict/statistics/at_glance/KeyTelecom.html.
6. Olinger, H., Britz, J., and Olivier, M. Western privacy and/or Ubuntu? Some critical comments on the influences in the forthcoming data privacy bill in South Africa. *The International Information & Library Review 39* (2007), 31–43.
7. The mother of invention. *The Economist* (Sept. 24, 2009), 8–12.

Seymour (Sy) Goodman (goodman@cc.gatech.edu) is Professor of International Affairs and Computing at Georgia Tech in Atlanta, GA.

Andrew Harris (harrisar@gatech.edu) is a researcher at Georgia Tech's Sam Nunn School of International Affairs in Atlanta, GA.

DOI:10.1145/1859204.1859216

Mark Smotherman and Dag Spicer

Historical Reflections

IBM's Single-Processor Supercomputer Efforts

Insights on the pioneering IBM Stretch and ACS projects.

Imagine a CPU designed to issue and execute up to seven instructions per clock cycle, with a clock rate 10 times faster than the reigning supercomputer. Imagine a design team of top experts in computer architecture, compilers, and computer engineering—including two future ACM A.M. Turing Award recipients, five future IBM Fellows, and five future National Academy of Engineering members. Imagine that this team explored advanced computer architecture ideas ranging from clustered microarchitecture to simultaneous multithreading.

Is this a description of the latest microprocessor or mainframe? No, this is the ACS-1 supercomputer design from more than 40 years ago.

In the 1950s and 1960s IBM undertook three major supercomputer projects: Stretch (1956–1961), the System/360 Model 90 series, and ACS (both 1961–1969). Each project produced significant advances in instruction-level parallelism, and each competed with supercomputers from other manufacturers: Univac LARC, Control Data Corporation (CDC) 6600, and CDC 6800/7600, respectively.

Of the three projects, the Model 90 series (91, 95, and 195) was the most successful and remains the most well known today, in particular for its out-of-order processing of floating-point operations. But over the past few years, new information about the other two efforts has surfaced, and many previously unseen documents are now physically collected and available online at the Computer History Museum in Mountain View, CA.

Stretch

Many histories recount that Stretch began in 1954 with efforts by Steve Dunwell.[2,11] Less known is that Gene Amdahl, designer of the IBM 704 scientific computer, was assigned to design Stretch.[10] Stretch was targeted at the needs of the Livermore and Los Alamos nuclear weapons laboratories, such as calculations for hydrodynamics and neutron diffusion. Unlike the vacuum-tube-based 704, which had identical machine and memory cycle times of 12 microseconds, the transistorized Stretch was expected to have a 100-nanosecond machine cycle time and a two-microsecond memory cycle time. In response to this logic/memory speed imbalance, Amdahl worked with John Backus to design an instruction look-ahead scheme they called asynchronous non-sequential (ANS) control.[10]

Both Amdahl and Dunwell participated in the sales pitch to Los Alamos in 1955. When Los Alamos contracted with IBM in 1956 to build Stretch, Dunwell was given control of the project and Amdahl left the company.

Dunwell recruited several new graduates including Fred Brooks and John Cocke. Harwood Kolsky, a physicist at Los Alamos, joined the project in 1957. Cocke and Kolsky developed a simulator that guided design decisions, particular-

The IBM Stretch supercomputer.

IMAGE COURTESY OF IBM

ly the details of the look-ahead. In its final form, branches in Stretch that could not be pre-executed were predicted, and instructions on the predicted path were speculatively executed—an astounding innovation for the late 1950s.[a]

Stretch pioneered the use of transistor technology within IBM, and the circuitry and memory modules designed for Stretch were used to bring the 7000-series of computers to market much earlier than would have been otherwise possible. Stretch pioneered many of the ideas that would later define the System/360 architecture including the 8-bit byte, a generalized interrupt system, memory protection for a multiprogrammed operating system, and standardized I/O interfaces.[3,4]

Stretch became the IBM 7030. A total of nine were built, including one as part of a special-purpose computer developed for NSA, codenamed Harvest. However Stretch did not live up to its initial performance goal of 60 to 100 times the performance of the 704, and IBM Chairman Tom Watson, Jr. announced a discount in proportion to this reduced performance and withdrew the 7030 from the market. Stretch was considered a commercial failure, and Dunwell was sent into internal exile at IBM for not alerting management to the developing performance problems. A number of years later, once Stretch's contributions were more apparent, Dunwell was recognized for his efforts and made an IBM Fellow.

Despite the failure, Kolsky and others urged management to continue work on high-end processors. Although Watson's enthusiasm was limited, he approved two projects named "X" and "Y" with goals of 10-to-20 and 100 times faster than Stretch, respectively.

Project X

Facing competition from the announcement of the CDC 6600 in 1963, Project X quickly evolved into the successful Model 90 series of the System/360.[6] Seventeen Model 91s and 95s were built. CDC sales and profits were impacted, and CDC sued IBM in December 1968 alleging unfair competition. The U.S. Department of Justice filed a similar antitrust suit against IBM one month later. CDC and IBM settled in 1973, and the government's antitrust action was dropped in 1982.

Although only one of these three projects was a commercial success, each significantly contributed to the computer industry.

Project Y

Project Y was led by Jack Bertram at IBM Research beginning in 1963. Bertram recruited Cocke, along with Fran Allen, Brian Randell, and Herb Schorr. Cocke proposed decoding multiple instructions per cycle to achieve an average execution rate of 1.5 instructions per cycle.[9] Cocke said this goal was a reaction to Amdahl, who after returning to IBM had postulated an upper limit of one instruction decode per cycle.[8]

In 1965, responding to the announcement of the CDC 6800 (later to become the 7600) and to Seymour Cray's success with a small isolated design team for the 6600, Watson expanded Project Y and relocated it to California near the company's San Jose disk facility. It was now called ACS (Advanced Computer Systems). The target customers were the same as for Stretch.

ACS-1

Amdahl was made an IBM Fellow in 1965, and Bob Evans encouraged him to informally oversee ACS. Although Amdahl argued strongly for System/360 compatibility, Schorr and other architects designed the ACS-1 around more robust floating-point formats of 48-bit single-precision and 96-bit double-

a The look-ahead supported recovery actions to handle branch mispredictions and provided for precise interrupts. Also, as in many current-day processors, complex instructions were broken into simpler parts ("elementalized" in Stretch terminology), with each part assigned a different look-ahead entry. However, this complexity was blamed for part of Stretch's performance problems, and speculative execution was not attempted in IBM mainframes until some 25 years later.

Calendar of Events

December 15–18
SIGGRAPH Asia 2010,
Seoul, Republic of Korea,
Contact: Hyeongseok Ko,
Email: ko@graphics.snu.ac.kr

December 17–18
First ACM Annual Symposium on Computing for Development,
London, United Kingdom,
Contact: Andy Dearden,
Email: a.m.dearden@shu.ac.uk

December 21–23
International Conference on Frontiers of Information Technology
Islamabad, Pakistan,
Contact: Muhammad Sarfaz,
Email: prof.sarfaz@gmail.com

January 4–8
The Third International Conference on Communication Systems and Networks,
Bangalore, India,
Contact: David B. Johnson,
Email: dbj@cs.rice.edu

January 5–8
Foundations of Genetic Algorithms XI,
Schwarzenberg, Austria,
Sponsored: SIGEVO,
Contact: Hans-Georg Beyer,
Email: hans-georg.beyer@fhy.at

January 17–20
The Thirteenth Australiasian Computing Education Conference,
Perth QLD Australia,
Contact: Michael de Raadt,
Email: deraadt@usq.edu.au

January 22–26
Fifth International Conference on Tangible, Embedded, and Embodied Interaction,
Funchal, Portugal,
Contact: Mark D. Gross,
Email: mdgross@cmu.edu

January 23–29
The 38th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages,
Austin, TX,
Sponsored: SIGPLAN,
Contact: Thomas J. Ball,
Email: tball@microsoft.com

precision as well as a much larger register set than System/360.[7] Amdahl was subsequently excluded and thereafter worked on his own competing design.

Average access time to memory in ACS-1 was reduced by using cache memory.[b] A new instruction prefetch scheme was designed by Ed Sussenguth. Pipeline disruption from branches was minimized by using multiple condition codes along with prepare-to-branch and predication schemes created by Cocke and Sussenguth. Compiler optimizations were viewed as critical to achieving high performance, and Allen and Cocke made significant contributions to program analysis and optimization.[1]

As for Stretch, detailed simulation was critical. Developed by Don Rozenberg and Lynn Conway, the simulator was used for documentation as well as validation and improvement of the design. While working through the simulation logic for multiple instruction decode and issue, Conway invented a general method to issue multiple out-of-order instructions per machine cycle that used an instruction queue to hold pending instructions. Named "Dynamic Instruction Scheduling,"[9] the scheme met the timing constraints and was quickly adopted.

Also like Stretch, advanced circuit technology was key to the performance goals. New circuits were developed with switching times in the nanosecond range, and new circuit packaging and cooling approaches were investigated.

In 1968, based on Amdahl's cost/performance arguments for his own design and the increasing cost projections for unique ACS-1 software, management decided to make ACS System/360 compatible and appointed Amdahl as director. The next year the project was canceled when Amdahl pushed for three different ACS-360 models. Instead, the cache-enhanced Model 195 was announced, and approximately 20 were built.

After ACS

Although the project fizzled, Silicon Valley benefited from the collection of talent assembled for the ACS project. Amdahl stayed in California and began his own company, with two dozen former ACS engineers joining him to build the Amdahl 470. Other ACS project members moved to IBM's San Jose disk drive facility or joined other companies in the Valley.

Cocke went back to the East coast, staying with IBM, and he later expressed regret that very little of the ACS-1 design was ever published.[5] Indeed, had the ACS-1 been built, its seven-issue, out-of-order design would have been the preeminent example of instruction-level parallelism. Instead, the combination of multiple instruction issue and out-of-order issue would not be implemented until 20 years later.

In discussing supercomputers in his autobiography, Watson blamed the "erratic" IBM efforts partly on his own "temper," and said he had come to view IBM's competition with CDC like General Motors' competition with Ferrari for building a high-performance sports car.[12]

In the 1970s the supercomputer industry pursued vector processing, with CDC building the unsuccessful STAR-100 in 1974 and Cray Research building the very successful Cray-1 in 1976. IBM announced a vector add-on for its mainframes in 1985, but the extension did not sell well. IBM funded Supercomputer Systems Incorporated between 1988 and 1993, but SSI went bankrupt before delivering its multiprocessor computer. More recently IBM built a number of successful large clusters, including Blue Gene and Roadrunner. In the June 2010 "Top500" Supercomputer list, four of the top 10 supercomputers in the world were built by IBM.

Impact

Although only one of these three projects was a commercial success, each significantly contributed to the computer industry. Fred Brooks has described the impact Stretch had on IBM, and especially on the System/360 architecture.[3] The Model 90 series influenced high-performance CPU design across the industry for decades. The impact of ACS was more indirect, through people such as Fran Allen, Gene Amdahl, John Cocke, Lynn Conway, and others. For example, Amdahl was able to recruit ACS engineers Bob Beall, Fred Buelow, and John Zasio to further develop high-speed ECL circuits and packaging for the Amdahl 470. Allen and Cocke disseminated their work on program optimization, and Cocke carried the idea of designing a computer in tandem with its compiler into later projects, one of which was the IBM 801 RISC. Conway went on to co-author a seminal book on VLSI design and attributes much of her insight into design processes to her ACS experience.

Historical Collections

The Computer History Museum collection contains more than 900 Stretch and ACS-related documents donated by Harwood Kolsky, and many of these are online at http://www.computerhistory.org/collections/ibmstretch/. These documents shed light on contributions to Stretch and ACS by Amdahl, Cocke, and others, as well as recording the design trade-offs made during the Stretch project and the studies made of Stretch performance problems. The Museum has several oral history interviews online at http://www.computerhistory.org/collections/oralhistories/ and has made recent talks available on its YouTube channel at http://www.youtube.com/user/ComputerHistory/. The Museum also features a new exhibit, part of which focuses exclusively on supercomputers, including the CDC 6600 and Cray-1. C

b Cache memory was a new concept at the time; the IBM S/360 Model 85 in 1969 was IBM's first commercial computer system to use cache.

References

1. Allen, F. The history of language processor technology in IBM. *IBM Journal of Research and Development 25*, 5 (Sept. 1981).
2. Bashe, C. et al. *IBM's Early Computers*. MIT Press, Cambridge, MA, 1986.
3. Brooks, F., Jr. Stretch-ing is great exercise—It gets you in shape to win. *IEEE Annals of the History of Computing 32*, 1 (Jan. 2010).
4. Buchholz, W. *Planning a Computer System: Project Stretch*. McGraw-Hill, Inc., Hightstown, NJ, 1962.
5. Cocke, J. The search for performance in scientific processors. *Commun. ACM 31*, 3 (Mar. 1988).
6. Pugh, E., Johnson, L. and Palmer, J. *IBM's 360 and Early 370 Systems*. MIT Press, Cambridge, MA, 1991.
7. Schorr, H. Design principles for a high-performance system. In *Proceedings of the Symposium on Computers and Automata* (New York, April 1971).
8. Shriver, B. and Capek, P. Just curious: An interview with John Cocke. *IEEE Computer 32*, 11 (Nov. 1999).
9. Smotherman, M. IBM Advanced Computing Systems (ACS); http://www.cs.clemson.edu/~mark/acs.html.
10. Smotherman, M. IBM Stretch (7030)—Aggressive Uniprocessor Parallelism; http://www.cs.clemson.edu/~mark/stretch.html.
11. Spicer, D. It's not easy being green (or "red"): The IBM stretch project. *Dr. Dobb's Journal*, July 21, 2001.
12. Watson, Jr., T. and Petre, P. *Father Son & Co*. Bantam, NY, 1990, 384.

Mark Smotherman (mark@cs.clemson.edu) is an associate professor in the School of Computing at Clemson University, SC.

Dag Spicer (spicer@computerhistory.org) is Senior Curator at the Computer History Museum in Mountain View, CA.

DOI:10.1145/1859204.1859217

Ann Quiroz Gates

Broadening Participation

The Role of Hispanic-Serving Institutions in Contributing to an Educated Work Force

Improving inclusiveness in computing education.

In order to thrive and even survive in the worldwide marketplace of ideas and innovation, the U.S. must aggressively meet the challenge of increasing the number of students who complete degrees in the fields of science, technology, engineering, and mathematics (STEM). It is critical for the economic and social health of the U.S. that a globally competitive STEM work force is maintained and the engagement of diverse individuals who can contribute to innovations and advancements in STEM areas is expanded. Although there has been an upturn in the past two years, computing fields have certainly experienced a significant decrease in the number of majors and graduates. Engaging large segments of society that have traditionally not been involved—students from underrepresented groups—is critical in addressing work force needs and innovation, especially in computing. One group in particular that is prime for greater inclusion in computing is Hispanics. Hispanics have the fastest growth rate among all groups in the U.S. (one in four newborns is Hispanic according to the Pew Foundation[5]), yet this group remains significantly underrepresented in STEM careers and in the number of graduates who obtain advanced degrees. Approximately 6.8% of the total bachelor's degrees awarded to citizens and permanent residents in 2000–2008 and 2.5% of the total doctoral degrees awarded in 1998–2007 in computer sciences went to Hispanics according to the *2009 NSF Women, Minorities, and Persons with Disabilities in Science and Engineering* report.[3] Adding to this is the fact that fewer younger adults are obtaining college degrees; the U.S. ranks only 10th in the percentage of the young adult population with college degrees according to the Lumina Foundation.

A Whittier College valedictory speaker during a Latino pre-graduation celebration.

PHOTOGRAPH BY REED SAXON / AP PHOTO

While it is imperative that we recruit and prepare a larger number of our youth for success in higher education, placing a focus on Hispanics is essential for closing the degree attainment gap particularly in STEM fields. While the numbers of Hispanics who continue on to baccalaureate programs and advanced studies are low, a number of social, educational, and environmental factors continue to

inhibit Hispanics from degree attainment. Recent survey data shows that Hispanic parents place high value on going to college; however, a large number of their children tend to drop out of school. This paradox may be due in large measure to poor socioeconomic conditions, the challenge of mastering the English language with little support, and the lack of role models. To make a difference, it is essential to have more Hispanic faculty in computing programs who can serve as exemplars and models at the community college and four-year institutions, to provide opportunities for development and growth, and to inform families about financial support structures for attending college.

The Role of Hispanic-Serving Institutions

With these issues in mind, we cannot ignore the role of Hispanic-serving institutions (HSIs) in educating Hispanics who become future leaders in the work force. Indeed, any effort to increase the number of Hispanics who attain STEM degrees will depend on the institutional capacity of community colleges and HSIs to educate and graduate Hispanics, in particular those who graduate in STEM fields. This claim is supported by a March 2010 report by Dowd et al.[2] from the Center for Urban Education that indicates a greater share of Hispanic students enrolled at HSIs earn degrees in key majors, such as computer science, mathematics, and engineering, than do their counterparts at non-HSIs.[a]

The report also states that Hispanic community college transfers who first earn associate's degrees have lower access to STEM bachelor's degrees at academically selective and private universities than their counterparts who do not earn an associate's degree prior to the bachelor's. Dowd's data shows that Hispanics, who are community college transfers and who first earn associate's degrees, have lower access to STEM bachelor's degrees at academically selective and private universities than their counterparts who do not earn an associate's degree prior to the bachelor's. On the other hand, transfer students were more likely to graduate from HSIs and from public four-year institutions, but they were less likely to graduate from academically selective institutions or from research universities. To make a difference, it is essential to support programs and initiatives that provide educational opportunities for Hispanics in STEM fields at HSIs and that target community college students and their successful transfer to four-year colleges.

The NSF Research Experiences for Undergraduates (REU) program has been a critical factor in providing students with research opportunities that motivate and prepare them for graduate studies. An SRI International evaluation of NSF support for undergraduate research opportunities found that students who participated in undergraduate research were twice as likely as those who did not do research to have pre-college expectations of obtaining a Ph.D.[6] In addition, participation in undergraduate research had strong positive effects on the students' understanding of the research process, confidence in their research-related abilities, and awareness of academic and career options in STEM. In regard to improving research opportunities, a common suggestion that arose from the study was the need for more effective faculty guidance and the key role of developing interpersonal, organizational, and research skills.

The Computing Alliance of HSIs, a consortium of 10 HSIs focused on the recruitment, retention, and advancement of Hispanics in Computing, has played an important role in involving students, in particular Hispanics, in research throughout the academic year using the Affinity Research Group (ARG) model and working with the students to apply for REU opportunities. ARG is focused on training faculty mentors on the ARG philosophy and how to structure in research groups the deliberate and intentional development of technical, team, and professional skills and knowledge required for research and collaborative work.

Arguments against having doctoral programs at minority-serving institutions (MSIs), such as those made by Richard Tapia in the March 2010 *Communications* Broadening Participation column,[7] often center on the low probability of individuals who do not graduate from top research institutions becoming a faculty members at those institutions and the supposed "lack of rigor" of MSI graduate programs, which make them less able to compete with the caliber of programs at top research institutions. Data reported in NSF 06-318 certainly supports the former argument,[1] although fewer than half of Hispanic faculty members who earned doctorates at top research institutions were employed at these institutions. The reasons are varied as are the reasons for graduates of top research institutions who choose to teach at

a National Center for Education Statistics study (NCES 2007-161) reports that only 10% of four-year institutions of higher education in the U.S. enroll the majority of Hispanic undergraduates.

FY2008 R&D Expenditures for Hispanic-serving institutions without a medical school.*

Ranking	Institution	R&D Expenditure (Dollars in thousands)	Percent Hispanic Enrollment**
32	NM State U. main campus	138,427	44.08
73	U. TX El Paso	48,906	74.94
93	CUNY The City C.	34,452	31.15
95	U. TX San Antonio	33,106	43.57
113	U. PR Rio Piedras Campus	22,662	99.94
118	U. PR Mayaguez Campus	20,763	100
129	U. CA, Merced	16,802	26.08
133	CA State U. Long Beach	14,971	25.44

* NSF/Division of Science Resources Statistics, *Academic Research and Development Expenditures: FY 2008*.[4]

** IPEDS Spring 2007 survey.

MSIs. One cannot assume that if individuals did not attend a top research institution, it is because they were not accepted, or that if graduates of top research institutions are not faculty members at a similar university, it is because they could not find a position. It may simply be nothing more than a conscious choice.

The second argument regarding "lack of rigor" at MSIs may reflect lack of awareness of the major transformation of a number of these universities in the past decade. As shown in the accompanying table, there are a number of HSIs that rank in the upper quarter of the rankings for FY 2008 R&D expenditures for all universities and colleges without a medical school. If we included MSIs, then four more institutions would be added to the table. The growth in research at HSIs can be seen when comparing the R&D expenditures for the institutions listed in the table over the last eight years. The total R&D expenditures for these institutions have risen steadily from $170,245,000 in 2001 to $258,773,000 in 2005, and to $330,089,000 in 2008.

There are a number of HSIs that are moving toward becoming national research universities through rigorous recruitment of excellent research faculty and establishment of strong collaborations with faculty and programs at top research institutions. The doctoral programs at many HSIs are built on the firm belief that broadening participation requires identifying students who have the capability to complete advanced studies, but who may lack the confidence to pursue such studies, and engaging such students in initiatives that prepare them to succeed in competitive research programs. It is imperative that we not equate skills development with the inhibition of a student's progress.

I would certainly agree that it is important to increase the number of Hispanic students who enter graduate studies at top research institutions, and faculty mentors should certainly encourage students to enter such programs; however, this effort will not make a large difference in the numbers. We cannot meet the challenge of increasing degree attainment from quality programs unless we continue to invest in research programs at MSIs and extend the number of universities and colleges that offer quality graduate education to a broader student population. We must ensure that students attending MSI programs have experiences that enhance their education, for example, by students spending a semester or summer at a top research institution, center, or laboratory; by MSI researchers establishing strong ties with top researchers at other institutions; and by students building effective networks and mentoring relationships.

We cannot continue to think in terms of the traditional models of education that have thrived on selectivity.

A Call to Action

Broadening participation is often discussed in terms of recruitment, retention, and advancement in academic programs. It also applies to broadening participation to those who serve in leadership roles such as members of national committees, policy debriefing committees, and advisory boards. While pedigree may be one measure to evaluate one's qualifications, there are other measures. Indeed, one's perspective, experiences, and accomplishments should serve as other important measures.

It is imperative that we work together to increase the percentage of U.S. residents, regardless of ethnicity, gender, or economic standing, who attain high-quality baccalaureate and advanced degrees. Democratization of higher education is essential to educate a broader base of citizens who are educated at a higher level. Toward this end, it is critical to support the efforts of HSIs and other MSIs to become national research universities. Work with the alliances, such as CAHSI and others that have formed under the CISE NSF Broadening Participation program, to build strong ties with MSI researchers and, as a result, their students.

Efforts such as these raise the quality of the university and graduates. It is through MSIs that we can effect change and meet the challenge of increased graduates in STEM. The counterargument to this position is that we need to encourage our "best" Hispanic students to attend the top universities in the country so that they have the credentials to assume leadership roles. It is irrefutable that students should study with the best minds in the country and establish the critical networks that are important for one's success. This does not mean, however, that some of the best minds cannot be found at MSIs and that there cannot be strong ties among researchers at different levels of research institutions that enhance the faculty's and students' research experiences. We cannot continue to think in terms of the traditional models of education that have thrived on selectivity. Now is the time to expand our thinking and to establish creative models for raising the quality of graduate education and providing accessibility to a larger group of people. Quite simply—and critically—the future of the U.S. depends on it. C

References

1. Burrelli, J. Academic institutions of minority faculty with S&E doctorates. *InfoBrief: Science Resources Statistics*, NSF06-318 (June 2006).
2. Dowd, A.C., Malcom, L.E., and Macias, E.E. *Improving Transfer Access to STEM Bachelor's Degrees at Hispanic-serving Institutions through the America COMPETES Act.* University of Southern California, Los Angeles, CA, 2010.
3. National Science Foundation, Division of Science Resources Statistics. *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2009*, NSF 09-305 Arlington, VA (Jan. 2009); http://www.nsf.gov/statistics/wmpd/.
4. National Science Foundation, Division of Science Resources Statistics. *Academic Research and Development Expenditures: Fiscal Year 2008.* Detailed Statistical Tables NSF 10-311. Arlington, VA (2010); http://www.nsf.gov/statistics/nsf10311/.
5. Pew Research Center. *Between Two Worlds, How Young Latinos Come of Age in America.* Pew Hispanic Center, Washington, D.C., 2009.
6. Russell, S. *Evaluation of NSF Support for Undergraduate Research Opportunities: Synthesis Report.* SRI Project P16346 (July 2006).
7. Tapia, R. Hiring and developing minority faculty at research universities. *Commun. ACM 53*, 3 (Mar. 2010).

Ann Quiroz Gates (agates@utep.edu) is associate vice president of research and sponsored projects at the University of Texas at El Paso and the recipient of the Anita Borg Institute for Women and Technology 2010 Social Impact Award.

DOI:10.1145/1859204.1859218

Peter J. Denning and Robert E. Kahn

The Profession of IT

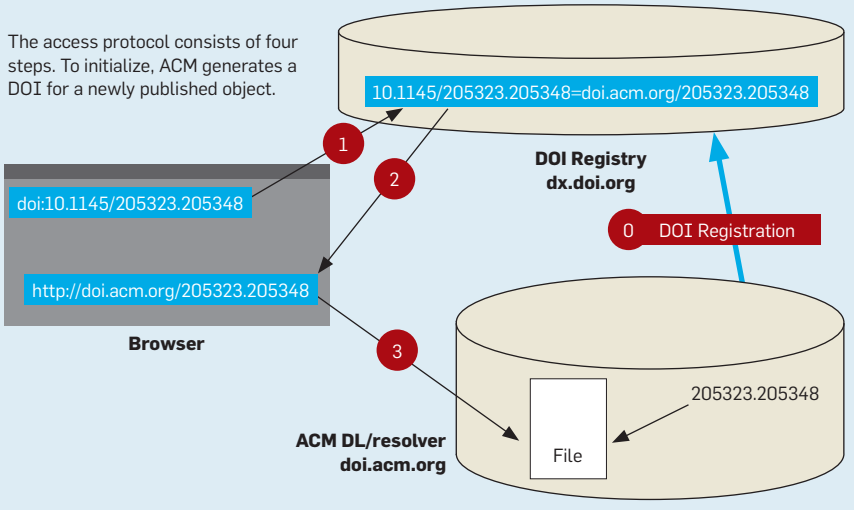
The Long Quest for Universal Information Access

Digital object repositories are on the cusp of resolving the long-standing problem of universal information access in the Internet.

Information sharing is an age-old objective. Always a challenge in the world of print documents and books, it has become truly daunting as music, movies, images, reports, designs, and other information entities are represented digitally and offered on the Internet. Numerous issues contribute to the complexity, such as file creation, formats, identifier systems (local and global), access controls, privacy controls, interoperability, searching, and rights management. The complexity is multiplied in the global Internet by the sheer amount of information available for access, the potential number of connections and reference chains, and jurisdictional issues. Reducing the complexity of information management in that universe has been a very long quest.

For the past 15 years a set of infrastructure technologies for "digital objects" has been gradually evolving in the Internet. They are now mature. We believe these technologies offer some real solutions to the conundrums of information sharing and access. We will summarize them and call attention to the significant user communities already employing them. We advocate that every IT professional become knowledgeable about these technologies and use them in the design of new systems and services.

Resolving a DOI to a URL or file name in the ACM Digital Library.



The DOI consists of ACM's unique number (10.1145) followed by a unique string chosen by ACM. ACM registers the DOI with the DOI registry (step 0). Thereafter a user can take the DOI from a citation and ask the registry to resolve it (step 1). The registry returns the URL of the object in the ACM Digital Library (step 2) . The ACM Digital Library resolver directs the access to the object specified by the DOI (step 3).

Early Attempts at Universal Information Access

Vannevar Bush is credited with the first visionary speculation about universal access to documents in 1945 ("As we may think," *Atlantic Monthly*). He proposed a hypothetical machine called Memex that stored documents on microfilm and allowed annotations and cross links. Many subsequent designers credit Bush with the inspirations for their work in computer networks.

The first among these designers was Ted Nelson, who, as early as 1960, proposed a networked system called Xanadu, which aimed at information sharing with a simple user interface. Nelson introduced topics such as hypertext, hyperlinks, automatic version management, automatic inclusion of referenced items, and small payments to authors for use of their materials.

In the middle 1960s, Doug Engelbart started the Augmentation Research Lab at SRI with the dream of amplifying collective intelligence through computer networks. Taking some inspirations from Bush and Nel-

son, he and his team developed NLS, the first working hypertext system with graphical user interface, mouse, and collaboration tools.

In the late 1980s, Tim Berners-Lee created the World Wide Web infrastructure to facilitate information access and as a potential means to implement many of Bush's, Nelson's, and Engelbart's ideas in the established worldwide Internet.

While these systems were major advances in the quest for universal information access, they left unsolved many of the difficult issues of information management noted earlier. Our objective here is to show that the Digital Object Architecture, which has been under development since the late 1980s at the Corporation for National Research Initiatives (CNRI) and is now reaching a tipping point, can provide the missing pieces of infrastructure and help us to attain universal information access.

Toward a Universal Address Space

A universal address space is the most fundamental element of a system of universal information access. A collection of ARPA-sponsored projects in the 1960s developed the first methods for doing this efficiently.

In a famous 1960 article "Man-computer symbiosis," J.C.R. Licklider expounded on the virtues of "intergalactic networks" and man-machine communications. Partly at his instigation, MIT's Project MAC undertook the construction of an operating system, Multics, which would be a "computer utility" that could dispense computing power and information access widely and cheaply. Other research organizations, such as IBM, UC Berkeley (Genie), and BBN (Tenex), were early developers of time-shared operating systems that could be networked.

Within the Multics system, information sharing was achieved by making virtual memory a common space accessible by every user on the system. The directory structure was an overlay that let users assign their own symbolic names to files; files themselves were addressed internally by their virtual addresses. Users never had to open or close files, or copy them from secondary storage to their workspaces; they simply referenced them as segments of address space.

A universal address space is the most fundamental element of a system of universal information access.

Jack Dennis, who helped design the Multics virtual memory, saw how to generalize it to allow dynamic sharing of a more general class of objects, not just files, and how to protect these objects from access not permitted by their owners. His "capability architecture" became the blueprint for object-oriented runtime systems.[2] That architecture inspired two commercial computing systems—Plessey 250 and IBM System 38—and two research projects—Cambridge CAP and CMU Hydra—that contributed much implementation knowledge. These projects all demonstrated that a large widely accessible address space would not only facilitate sharing, but it could be implemented efficiently on a single machine with a large shared file system. The capability architecture also provided a clean way of managing access and controlling rights by channeling all object references through a reference monitor protocol.[3]

The capability architecture easily generalized to homogeneous networks all running the same operating system. Unfortunately, it did not generalize well to the heterogeneous systems making up the Internet.

In 1970, the ARPANET expanded the universe of connected systems beyond a single machine to hundreds of time-shared computers. In the U.S. and around the world, other packet networks were developed in parallel with the ARPANET, including, in particular, a packet satellite network, a ground radio packet network, and various local area networks, such as token rings and Ethernet.

The Internet, introduced in 1973 by Bob Kahn and Vint Cerf, is a global information system composed of many networks and computational resources all embedded within a single large address space based on the Internet Protocol (IP) addresses. The domain name system, introduced in 1983, maps domain names to IP addresses, making it much easier for users to address computers.

Information sharing was accomplished in the Internet in various ways. In the early 1970s, the primary means were text-based email and file transfer, followed later by email attachments. A few research communities pioneered with collections of reports and software collections hosted on specific machines. Professional associations such as ACM built digital libraries to share their entire published literature. Services to back up files and provide drop points for transferring files became common.

In 1989, the Web became available as an architecture overlaid on the Internet to facilitate information sharing. The Web protocols automatically invoked a file transfer protocol with a single mouse click on a hyperlink. Hyperlinks used Uniform Resource Locators (URLs) to name objects; a URL consisted of a host name concatenated with the pathname of a file on the host file system.

The Web was an important step toward universal information sharing but it is far from perfect. Among other things, URLs are not persistent, the same URL can be reused for different purposes, and users frequently encounter broken links because files have been moved to new locations. Moreover, there are no rights-management protocols, and the only access controls are those for file access on target machines.

The Digital Object Architecture

Universal information access was not achieved in the Internet because none of the protocols was designed relative to information-sharing architecture principles. In the 1980s, prior to the development of the Web, as part of its work on digital libraries, CNRI started designing a system for enabling mobile programs (called "Knowbots") to carry out information-access tasks in a network environment. This led to the later formulation by CNRI of the Digital Object Architecture (DOA),[1,4,5] which culled out and unified four key

principles from the past projects on information access:

▸ Any unit of information represented in digital form may be structured as a digital object (DO) for access (with suitable controls) within the Internet. DOs may include digitized versions of text files, sounds, images, contracts and photos, as well as information embedded in RFID devices, chip designs, simulations, or genome codes. The structure of a DO, including its metadata, is machine and platform independent.

▸ Every DO has a unique persistent identifier, called a "handle," or generically, a "digital object identifier," that can distinguish a DO (or separately identified parts of it) from every other object, present, past, or future. Handles consist of a unique prefix allotted to an entity (such as a publisher or individual) followed by a string of symbols chosen by the entity. The "resolution" system maps handles to state information that includes location, authentication, rights specifications, allowed operations, and object attributes.

▸ DOs can be stored in DO Repositories, which are searchable systems that offer continuous access to objects over long time intervals that span technology generations.

▸ Accesses to an instance of DO Repository are made via a standard DO protocol that restricts actions to those consistent with an object's state information.

The primary components of CNRI's DOA design are the Handle System, DO Repositories, and DO Registries:

▸ The Handle System allots prefixes to registered administrators of local handle services and provides resolution services for their digital object identifiers. This system has been available as a service on the Internet since the middle 1990s and is highly reliable. It makes use of existing Internet protocols, which do not need redesign. Handle services also support domain name resolution for backward compatibility.

▸ The DO repositories use standard storage systems. They provide digital object management services with a standard protocol called digital object protocol (DOP). The DO repositories also support HTTP and DOP-over-TLS, a secure socket layer (SSL) service.

▸ The DO registries allow users to reference, federate, and otherwise manage collections across multiple repositories and allow for protected access to such information including completely private information, sharing within designated groups, and full public access.

The Web was an important step toward universal information sharing but it is far from perfect.

Examples of DOA in Use

The figure on the first page of this column shows the essence of a resolution of a DOI to a URL or a file name, using the ACM Digital Library as an example.

The International DOI Foundation (IDF) is a non-profit organization that administers DOIs for a variety of organizations, mostly publishers. The IDF has trademarked the DOI; and it uses the Handle System technology for resolution. One of the IDF Registration Agents (RAs), DataCite, manages large scientific data sets using DOIs. The largest of the IDF RAs, CrossRef, manages metadata on behalf of a large segment of the publishing industry.

The U.S. Library of Congress uses the Handle System to identify large parts of its collections. The U.S. Department of Defense (http://www.adlnet.gov) relies on the Handle System and DOI Registry and Repository to manage distributed learning material. The European Persistent Identifier Consortium (EPIC) (http://www.pidconsortium.eu) provides identifier services to the European research community. People in the legal community are implementing the DOA for wills, deeds to real property, bills of lading, and other legal instruments.

Conclusion

The quest for universal information access in networks began around 1960 and over the years yielded a set of principles to fully support universal information access. These principles include unique, persistent identifiers, protocols that map identifiers to objects (including their metadata), protocols for enforcing access and use rights, and repositories that hold objects for indefinite periods spanning technology generations. These principles offer a possible solution to universal information access and an infrastructure for more general information management.

The Digital Object Architecture was designed to exploit these principles without changing existing Internet protocols. The architecture is now widely used by publishers, digital libraries, government agencies, and many others to manage their collections and offer them for access over the Internet.

The Digital Object Architecture offers computing professionals an emerging infrastructure for managing information that goes a long way toward universal information access as well as system interoperability in the Internet. We advocate that every computing professional become familiar with these technologies and use them for new systems and applications. The Internet sites doregistry.org, dorepository.org, and handle.net contain tutorial materials and are a good place to start learning how these technologies work. C

References

1. Corporation for National Research Initiatives. *A Brief Overview of the Digital Object Architecture and its Application to Identification, Discovery, Resolution and Access to Information in Digital Form* (June 2010); http://www.cnri.reston.va.us/papers/Digital_Object_Architecture_Brief_Overview.pdf
2. Dennis, J.B.and Van Horn, E. Programming semantics for multiprogrammed computations. *Commun. ACM 9*, 3 (Mar. 1966), 143–155.
3. Graham, G.S. and Denning, P. Protection: Principles and practice. *AFIPS SJCC Conference* (May 1972), 417–429. DOI: 10.1145/1478873.1478928.
4. Kahn, R.E. and Lyons, P. Representing value as digital objects: A discussion of transferability and anonymity. *Journal of Telecommunications and High Technology Law 5* (2006).
5. Kahn, R.E. and Wilensky, R. A framework for distributed digital object services. *International Journal on Digital Libraries 6*, 2 (2006). DOI: 10.1007/s00799-005-0128-x. (First made available on the Internet in 1995 and reprinted in 2006 as part of a collection of seminal papers on digital libraries).

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

Robert E. Kahn (rkahn@cnri.reston.va.us) is President and CEO of the Corporation for National Research Initiatives and a recipient of the ACM A.M.Turing Award.

DOI:10.1145/1859204.1859219

George V. Neville-Neil

Article development led by acmqueue
queue.acm.org

Kode Vicious
Literate Coding

Spelling and grammar do matter.

Dear KV,

Do spelling and punctuation really matter in code? I'm a freshman studying computer science at college, and one of my professors actually takes points off of our programs if we spell things incorrectly or make grammar mistakes in our comments. This is completely unfair. Programmers aren't English majors and we shouldn't be expected to write like them.

Miss Spelled

Dear Miss Spelled,

Who is your professor and can I please send him or her some flowers? Normally I find myself wanting to send people black roses, or a horse's head wrapped in paper, but your letter makes me want to send whoever is teaching you to code a big bouquet of fresh flowers, and a card, and maybe one of those ridiculous balloons that people now send as well.

While it is true that "programmers aren't English majors," there are days—many, many days—that I wish they were, or that they knew one, even tangentially, and offered to help with their science or math homework in return for some help making themselves better understood. The amount of code I have had to beat my head against because the original author did not make his or her intent clear in the comments, when there were comments, is legion, and all that head beating has given me many headaches. It cheers me to know someone else out there is beating your head for me, so maybe, someday, my headaches will go away. I also think the voices might stop at that point, too.

Clear code is actually more aligned with clear writing than most people are willing to admit. Many programmers seem to have gone into their field to avoid subjects such as English or anything that would require communicating with other people rather than machines. That is unfortunate, because code is a form of human communication.

Although a computer will execute your program, it will not execute it in the exact form in which you wrote it, and it definitely does not care that your variables are misspelled or that

your comments do not make it clear who did what to whom in the code that follows them. Code is not just for computers; it is for other people as well. If the next person who reads your code—in your case your professor, but after school it will be your workmates—cannot make heads or tails of it because you failed to write in a clear manner, then that is sloppy work on your part and deserves poor marks.

There are many defects that make reading code difficult: from poor structure, to poor formatting, to poor spelling. Attention to detail is the hallmark of all good coders, and that's not just detail in terms of the executing lines of code; it's all of the details, including the words you use and how you use them.

If you are a really poor speller, then run a spell checker over your code before you submit it; and if English is not your first language, then ask one of your classmates to read your comments and tell you if they make sense.

Remember, whether you like them or not, other people are going to read your code, and you owe it to them to make what you wrote easy to read.

KV

Dear KV,

I think I have now come across possibly the most stubborn and recalcitrant programmer I have ever met. He may even rival you in ability to complain about stupid things he encounters. While he's an amazing coder who can produce clear and elegant programs, he has one odd blind spot: he hates debuggers. I don't understand why, and asking him only gets you either a long diatribe or strange mumbling, so I've stopped asking. His preferred form of debugging is via `print` statements, no matter what language he is working in. While I understand that a debugger is not a replacement for writing good code, it's a tool I wouldn't really want to be without. Why would anyone choose `print()` over a debugger?

Bugged and Debugged

Dear Bugged,

I never understand why people who read my columns think of me as stubborn or recalcitrant. Despite my pen name, I am nothing but sweetness and light, and I never complain—I simply point things out.

Clear code is actually more aligned with clear writing than most people are willing to admit.

With that admittedly ridiculous denial out of the way, let's talk about debuggers and people who do and don't use them. Yes, there are people who find debuggers to be an inferior tool and who prefer to use in-program logging, or `printf`, statements to find out where their program is going wrong. I suspect you have also heard about people who cut their own quills and make their own ink for writing.

A debugger is a tool, and if the tool works for you, then you should use it. I see no reason to shun or depend solely on a single tool to help me correct my programs—not that KV ever has bugs in his code.

All that being said, there are many places where a `print` statement is what you actually want. It's important to remember that the debugger interferes with the program it is debugging and can lead to problems with observation. Armchair physicists love to talk about Schrödinger's cat, the thought experiment wherein observing an experiment has the effect of changing the outcome of the experiment. A debugger can have the same effect, and can be far more pernicious than the simple `print` statement. When you're using a debugger on a program, the debugger is introducing overhead, stopping and starting your program using signals, and generally manipulating your program like a sock puppet. While sock puppets are amusing, they're not subtle, and so subtle bugs may not show themselves when you're using the debugger.

A `print` statement, on the other hand, is a simple way to find out if something has gone drastically wrong with your code, and this is why good coders add logging to their systems almost from the start. I've even seen someone use `print` statements to debug multiprocessor locking in an operating system, by printing messages to two different consoles, one tied to each of the CPUs.

The biggest problems I see with using `print` statements for debugging are that they are often left active when they should not be, and in many cases, adding the `print` statements changes the layout of the code in memory. If you're looking for any kind of complex bug—for example, a memory smash, threading, or timing issue—then the humble `print` is more likely to cover up the bug than to help you find it. In these cases, you're better off with a good debugger and hardware that supports watchpoints. A watchpoint is a way to see if a memory location changes at runtime, and it is the best way to find memory smashes in your code—that is, bugs where you're accidentally overwriting memory.

As I'm sure you expected, KV comes down on the side of having many tools—from rough clubs and machetes to finely honed scalpels—for killing bugs.

KV

Related articles on queue.acm.org

A Conversation with Steve Bourne, Eric Allman, and Bryan Cantrill
http://queue.acm.org/detail.cfm?id=1454460

Kode Vicious: Gettin' Your Head Straight
George V. Neville-Neil
http://queue.acm.org/detail.cfm?id=1281885

George V. Neville-Neil (kv@acm.org) is the proprietor of Neville-Neil Consulting and a member of the ACM *Queue* editorial board. He works on networking and operating systems code for fun and profit, teaches courses on various programming-related subjects, and encourages your comments, quips, and code snips pertaining to his *Communications* column.

DOI:10.1145/1859204.1859220

Francine Berman

Viewpoint
We Need a Research Data Census

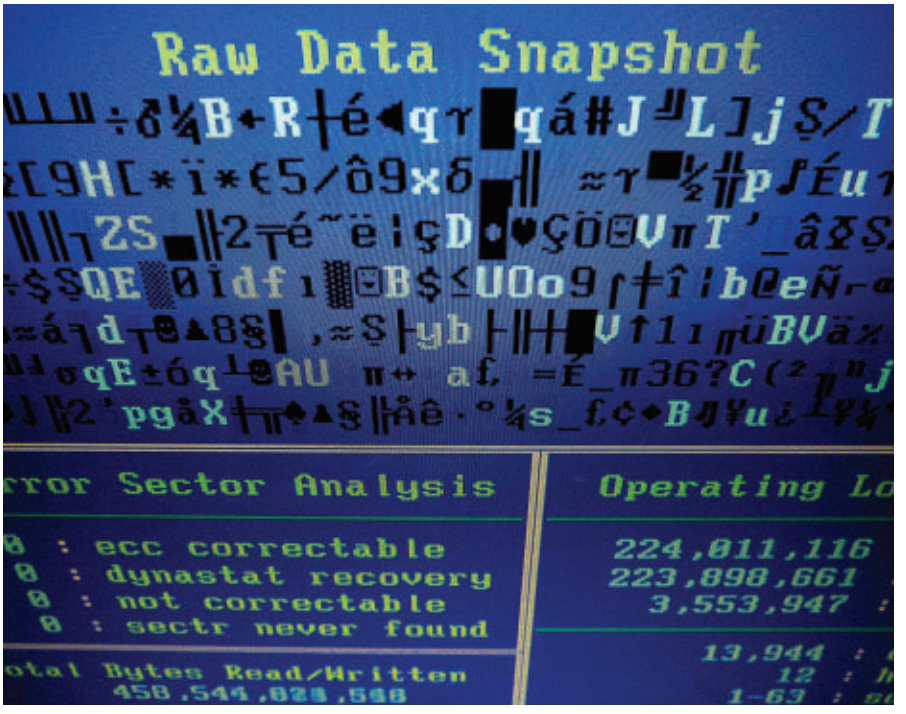
The increasing volume of research data highlights the need for reliable, cost-effective data storage and preservation at the national scale.

THIS PAST YEAR was a census year in the U.S. We responded to arguably the most long-lived and broad-based gathering of domiciliary information about the American public anywhere. U.S. Census data, collected every decade, provides a detailed picture of how many of us there are, where we live, and how we're distributed by age, gender, household, ethnic diversity, and other characteristics.

The Census (http://2010.census.gov/2010census/index.php) provides an evidence-based snapshot of America. This important information is publicly available and used in a variety of ways—to guide in the planning of senior centers, schools, bridges, and emergency services, to make assessments informed by societal trends and attributes, and to make predictions about future social and economic needs. The Census is particularly valuable as a planning tool in the building of physical infrastructure, as the distribution and characteristics of the population drive the development of hospitals, public works projects, and other essential facilities and services.

Given the role and importance of the Census in the physical world, it is useful to ask what provides an analogous evidence-based and publicly available snapshot of the "inhabitants" of the Digital World—our digital data.

What do we know about our data? How much is there? Where does it reside? What are its characteristics? Good



"top-down" methodological estimates of these questions have come from the reports on the increasing deluge of digital information developed by the IDC (http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf), by Bohn and Short (http://hmi.ucsd.edu/pdf/HMI_2009_ConsumerReport_Dec9_2009.pdf), and (some time ago) by Lyman and Varian (http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf). These provide intriguing, analytically derived bounds of the Digital World.

However, to make economic decisions that can drive the cost-effective development and deployment of the cyberinfrastructure needed to support long-lived digital data, we need more resolution. This is particularly important in the research arena, where federal R&D agencies apportion funding between the competing priorities of conducting basic research, and creating and supporting the cyberinfrastructure that enables that research. Just as the U.S. Census drives planning for infrastructure in the physical world, a Research Data Census would inform

PHOTOGRAPH BY PAUL DINEEN

cost-effective planning for stewardship of federally funded, shared cyberinfrastructure in the Digital World.

A Census for Research Data

The 10 questions on the 2010 U.S. Census form are well defined and provide basic information. There are many things not addressed in the Census—educational level of the population, for example. Similarly, an effective Research Data Census should provide basic information about the research data generated from federal funding. It should help us design, develop, and identify appropriately sized and outfitted storage, repositories, and services in the Digital World. It should provide a quantitative snapshot of the research data landscape at a given point in time, exposing key characteristics, such as:

▸ **Number and size distribution of federally funded research data sets.** How many research data sets generated by federal funding are less than a terabyte (that is, host-able on a researcher's hard drive), between 1 and 100 terabytes (perhaps host-able at a university repository), between 100 terabytes and a petabyte (perhaps requiring a larger-scale shared archive), more than a petabyte? What is their distribution?

▸ **Type and area distribution of federally funded research data sets.** What percentage of the U.S. federally funded research data is text, video, audio, and so forth? How much digital research data is generated within specific research areas (as categorized by NSF Directorates, NIH institutes, and other groups)?

▸ **Needs for preservation.** How much federally funded digital data must be retained by policy or regulation (HIPAA, OMB A-110, and so forth) for up to 1 year, 1–3 years, 3–5 years, 5–10 years, more than 10 years?

▸ **Common services and tools.** What categories of services and tools (gene sequence analysis, data visualization, mosaicing, and so forth) are used in conjunction with federally funded research data sets?

Note that basic questions along these lines will *not* provide a complete picture of our data. They do not differentiate between derived data and source data, for example; nor do they

provide comprehensive information about the necessary data systems and environments required to support data.

A Research Data Census *will* provide some specifics critical to cost-effective planning for stewardship of federally funded research data, however, and it will allow us to infer some key requirements for data cyberinfrastructure. In particular, a Research Data Census could help inform:

- **Useful estimates of the storage capacity required for data stewardship, and a lower bound on data that must be preserved for future timeframes.** Data required by regulation or policy to be preserved is a lower bound on valued preservation-worthy research data—additional data sets will need to be preserved for research progress (for example, National Virtual Observatory data sets).
- **The types of data services most important for research efforts.** Knowing the most common types of useful services and tools can help drive academic and commercial efforts.
- **Estimates of the size, training, and skill sets that will be needed for today's and tomorrow's data work force.**

Getting It Done

A Data Census sounds like a big job and it is, however there is potential to use existing mechanisms to help gather the needed information efficiently. We already provide annual and final reports to federal R&D agencies to describe the results of sponsored research. One could imagine a straightforward addition to annual reporting vehicles and/or sites such as grants.gov to collect this information (preferably electronically). Although U.S. Census information is gathered every 10 years, the Research Data Census would require frequent updating in order to provide useful information for planning purposes about our dynamically changing data landscape. The right periodicity for reporting is a topic for discussion, but an annual update probably provides the best resolution for the purpose of tracking trends.

Note also that there is real complexity in doing an effective Data Census: much of our data is generated from collaborative research, which can cross institutional, agency, and national boundaries. The Data Census reporting mechanisms must take this into account to produce relatively accurate counts. Data sets are often replicated for preservation purposes—do we count the data in all copies (all of which require storage), or do we count only the non-replicated data? (It is interesting to note that the U.S. Census has a related problem and covers it as question 10: "Does person 1 sometimes live or stay somewhere else?" If yes, check all that apply....). As with any survey, careful design is critical in order to ensure the results are accurate and useful as the basis for making predictions and tracking trends.

An effective Research Data Census should provide a quantitative snapshot of the research data landscape at a given point in time.

Using the Research Data Census to Create Effective Data Stewardship

An important outcome of the Research Data Census would be evidence-based information on the amount of data in the research community that must be preserved over time. This would help in understanding and meeting our needs for archival services and community repositories.

Such information can help cut data management and preservation problems down to size. Knowing that data valued by a particular community is typically of a certain type, a certain size, and/or needed over a certain timeframe, can help the community plan for the effective stewardship of that data. For example, accurate estimates of the digital data emanating from the Large Hadron Collider at CERN have been instrumental in the development of a data analysis and management plan for the High Energy Physics community.

It is likely that some of the capacity needed for stewardship of research data will come from university libraries reinventing themselves to address 21st century information needs; some of the capacity may come from the commercial sector, which has responded to emerging needs for digital storage and preservation through the development of commercial services. In some cases, the federal government will take on the stewardship responsibilities for research data (for example, the NIST Science Reference Data). It is clear that the size, privacy, longevity, preservation, access, and other requirements for research data preclude a "one-size-fits-all" approach to creation of supporting data cyberinfrastructure. It is also true that no one sector will be able to take on the responsibility for stewardship of all research data. A national strategic partnership spanning distinct sectors and stakeholder communities is needed to effectively address the capacity, infrastructure, preservation, and privacy issues associated with the growing deluge of research data. The development of a Research Data Census can provide critical information for more effectively developing this partnership.

No Time Like the Present

The 2010 requirement for a data management plan at the National Science Foundation (http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928&org=NSF) joins existing requirements for data sharing and management at NIH and elsewhere. Such requirements expand community awareness about responsible digital data stewardship and will exacerbate the emerging need for reliable, cost-effective data storage and preservation at the national scale.

A Research Data Census will provide a foundation for estimating the data cyberinfrastructure required for strategic stewardship. It can lay the groundwork today for access to our most valuable digital research assets tomorrow, and the new discoveries and innovation they drive. C

Francine Berman (bermaf@rpi.edu) is Vice President for Research at Rensselaer Polytechnic Institute, the former director of the San Diego Supercomputer Center, and the co-chair of the Blue Ribbon Task Force for Sustainable Digital Preservation and Access (http://brtf.sdsc.edu).

DOI:10.1145/1859204.1859224

Article development led by acmqueue
queue.acm.org

Pixar's president Ed Catmull sits down with Stanford professor (and former Pixar-ian) Pat Hanrahan to reflect on the blending of art and technology.

A Conversation with Ed Catmull

WITH THE RELEASE of *Toy Story* in 1995, Pixar Animation Studios President Ed Catmull achieved a lifelong goal: to make the world's first feature-length, fully computer-generated movie. It was the culmination of 20 years of work, beginning at the legendary University of Utah computer graphics program in the early 1970s, with important stops along the way at the New York Institute of Technology, Lucasfilm, and finally Pixar, which he cofounded with Steve Jobs, Alvy Ray Smith, and John Lasseter in 1986. Since then, Pixar has become a household name, and Catmull's original dream has extended into a string of successful computer-animated movies. Each stage in his storied career presented new challenges, and on the other side of them, new lessons. Here, Catmull shares some of the insights he has gained over the past 40 years, from the best way to model curved surfaces to how art and science interact at Pixar.

Interviewing Catmull is Stanford computer graphics professor Pat Hanrahan, a former Pixar employee who worked with Catmull on Pixar's acclaimed RenderMan rendering software, for which they share a Scientific and Engineering Academy Award. Hanrahan's current research at Stanford focuses on visualization, image synthesis, virtual worlds, and graphics systems and architectures. He was elected to the National Academy of Engineering and the American Academy of Arts and Sciences and is a member of *ACM Queue*'s Editorial Advisory Board.

PAT HANRAHAN: You're one of those lucky guys who got to go to the University of Utah during what was probably one of the most creative times in computing. All these great people were there. You were advised by Ivan Sutherland; Jim Clark was there, as was Alan Kay, John Warnock, and many other incredible people. What was it like? Did you all hang out together and work together?

ED CATMULL: We were all quite close. The program was funded by ARPA (Advanced Research Projects Agency), which had an enlightened approach, and our offices were close together, so it was a great social environment.

Dave Evans was the chairman of the department and Ivan was teaching, but their company, Evans and Sutherland, took all their excess time. The students were pretty much independent, which I took as a real positive in that the students *had* to do something on their own. We were expected to create original work. We were at the frontier, and our job was to expand it. They basically said, "You can consult with us every once in a while, and we'll check in with you, but we're off running this company."

I thought that worked great! It set up this environment of supportive, collegial work with each other.

PHOTOGRAPH BY TOM UPTON

HANRAHAN: It seems that you published papers on every area of computer graphics during that era. Your thesis—and I don't even know if you ever published a lot of what was in there—described z-buffering, texture-mapping algorithms, and methods for the display of bicubic surfaces. You wrote papers on hidden-surface algorithms, anti-aliasing, and, of course, computer animation and geometrical modeling. Your interests seem so wide ranging—not the way typical grad students today approach research, where they'll home in on one particular subtopic and drill down to bedrock on it.

CATMULL: I guess I didn't know any better.

HANRAHAN: Did you have a favorite from all of that work? What do you think was your most inspirational thought during that period?

CATMULL: For me, the challenge was to figure out how to do *real* curved surfaces, because other than quadric surfaces, which were way too limited, everything was made up of polygons. So, my first idea was to find a way to actually bend polygons to have them look right.

HANRAHAN: How do you bend a polygon?

CATMULL: Well, it was kind of ad hoc. And it largely worked, but as you might imagine, if you don't have a well-defined surface, then you're going to come up with cases that break. So, I figured out how to work with B-spline surfaces, but the difficulty was it would take way too long to render a picture. At the time, I thought my most ingenious idea was a new method for subdividing a surface. It was the equivalent of a finite difference equation, except that it split the curve in half with each iteration. Think of a line: with a difference equation, four adds gets you the next point in a cubic curve; every four adds gets you a new point. I came up with a method such that every four adds got you the point in the middle of the curve. It turns out that for doing subdivision surfaces, it's really fast.

HANRAHAN: Can it be made recursive?

CATMULL: It was recursive. I thought this was my neatest idea at the time, and to my knowledge, it actually was. It's an original contribution for difference equations. But as computers got faster, it just wasn't the main problem. I implemented it, and sure enough, it was fast; but it still took 45 minutes to make pictures on the PDP-10.

HANRAHAN: Evans and Sutherland Corporation placed heavy emphasis on real-time graphics, but you were willing to step back a little and say, "I know there's this emphasis on making things efficient, but we should explore this frontier of non-real-time stuff, and see what's possible—not just what we can do in real time at the time."

CATMULL: It's true there was a division, but the breadth of the support from Dave and Ivan encompassed both approaches. At the time, I wanted to develop the technology so we could make motion pictures. I had a very clear goal.

HANRAHAN: Was your goal different from the rest of them?

CATMULL: Yes, but that was perfectly fine. In fact, Ivan at one point started a film company called the Electric Picture Company. He was going to hire Gary Demos and me, but they couldn't get the funding, so it fell apart. Ivan had an interest in pushing the direction of motion pictures, and he knew that was my drive, too. I was working on the good-looking pictures, and they were working on interactivity.

HANRAHAN: When I first got interested in graphics in grad school, I heard about this quest to make a full-length computer-generated picture. At the time I was very interested in artificial intelligence, which has this idea of a Turing test and emulating the mind. I thought the idea of making a computer-generated picture was a prelim to, or at least as complicated as, modeling the human mind, because you would have to model this whole virtual world, and you would have to have people in that world—and if the virtual world and the people in it didn't seem intelligent, then that world would not pass the Turing test and therefore wouldn't seem plausible.

I guess I was savvy enough to think we weren't actually going to be able to model human intelligence in my lifetime. So, one of the reasons I was interested in graphics is I thought it had a good long-term career potential. I never thought when I entered the field that by the time I died we would have made a fully computer-generated picture, but I thought it would be great fun, and eventually it would happen. Did you ever have thoughts like that?

CATMULL: Oh yes, but when I graduated my goal was not as lofty as emulating all of reality; it was to make an animated film. That was more achievable, and I thought that it would take 10 years. This was 1974, so I thought by 1984, we might be able to do it. I was off by a factor of two: it took 20 years. I remember giving talks and saying at the time, "Look at the table in front of you. Nobody has been able to make a picture that comes anywhere close to capturing the complexity of even just that table in front of you." As we started to get close to that, it stopped being a meaningful thing to say. And, of course, now we're way past that.

We believed that achieving the appearance of reality was a great technical goal—not because we were trying to emulate reality, but because doing it is so hard that it would help drive us forward. That is, in fact, what happened. We were trying to match the physics of the real world, and in doing that we finally reached the point where we can create convincingly realistic images. Reality was a great goal for a while. Now we have non-photorealistic rendering goals and other things like that that have supplemented it. For a number of years, animation and matching reality were very useful goals, but I never thought of them as the ultimate goal.

We were also fairly good at analyzing how much compute power it would take, and this was at a time when others were buying Cray computers. We were at Lucasfilm at the time, and the feeling was that if anybody could afford a Cray computer it would be Lucasfilm. But from our point of view it was nuts, because, we asked, "How much compute power will it take to do a whole movie?" The answer was that it would take 100 Cray-1s. We realized we were so far away that we shouldn't even waste time being jealous. There were other things we had to do.

HANRAHAN: That's an interesting story because you had this 10-year vision and you didn't try to do it too soon. If something is too far in the future, typically you can't enumerate everything that must be done to achieve your goal; but you seem to have figured out the steps, and you kept building toward that ultimate goal.

CATMULL: I always believed that you need to look at the steps—we had to consider what the computing would provide, the economics, and the software solutions.

And if you look at the underlying infrastructure, even at that time we all knew Moore's Law and that it was going to continue into the foreseeable future. We knew there were a lot of things we didn't know how to do, and we could list what they were: modeling, animation, and simulation. Back then, those were the clear problems we had in front of us. We didn't want to waste money on a Cray, because we hadn't solved these other problems.

HANRAHAN: That was a very brilliant analysis. But how did you get the funding? These days it would be really hard to convince somebody to fund me for 10 years. Yet, you've worked with various people—first Alexander Schure, then George Lucas, and then Steve Jobs—and they all seemed willing to invest in long-term development.

CATMULL: Interestingly enough, when I graduated from Utah, I tried to go to another university, but I couldn't find one that would buy into that long-term plan. Of course, I had just come out

of a place where they did do that, and I always looked at ARPA at that time as being a spectacularly successful example of governmental policy moving things in the right direction. It had very low bureaucracy and trusted that if you give funding to smart people, some interesting things will happen. I came up out of that environment, and to this day I believe it's a great thing to do. That doesn't mean you won't have some failures or some abuses along the way, but that model of funding universities was spectacularly successful.

Unfortunately that wasn't the way it worked at the rest of the schools I was applying to, so I originally got a job doing CAD work at Applicon. Then Alex Schure [founder of the New York Institute of Technology] came along, and he wanted to invest in animation and make me head of the computer graphics department.

He didn't have all of the pieces necessary to do it, but he was the only person willing to invest in it. We had this remarkable group of software people coming to New York Tech, and Alex was essentially supporting them, but the technical people there knew that an element was missing: they didn't have the artists or the other components of filmmaking. Alex didn't understand that. He thought we were the filmmakers, and that rather than being part of a larger thing, that we were the solution. Unfortunately, he never got full credit for what he did because of that little bit of a blind spot on his part. He certainly made a lot happen for which he hasn't gotten a lot of credit.

Eventually, I moved on to Lucasfilm, where a very interesting thing happened, which I don't think people quite understand. At the time Lucasfilm was making the second *Star Wars* film, and the people at ILM [Industrial Light and Magic, Lucasfilm's special effects division] were the best guys in the world at special effects. So George [Lucas] took me to them and said, "OK, we're going to do some computer graphics." These guys were very friendly and very open, but it was extremely clear that what I was doing was absolutely irrelevant to what they were doing.

HANRAHAN: They didn't get it?

CATMULL: They didn't think it was relevant. In their minds, we were working on computer-generated images—and for them, what was a computer-generated image? What was an image they saw on a CRT? It was television.

ON THE EARLY DAYS:

We believed that achieving the appearance of reality was a great technical goal—not because we were trying to emulate reality, but because doing it is so hard it would help drive us forward. That is, in fact, what happened. We finally reached the point where we can create convincingly realistic images. Reality was a great goal for a while.

Even if you made good television, it looked crappy by their standards. However, from their point of view, it was George's decision and he could do with his money what he wanted. It wasn't as though he was taking anything away from them—it's just that computer graphics was not relevant to what they were doing on that film.

I look back at this and see that what we had—and this is very important—was *protection*. When you're doing something new, it's actually fairly fragile. People don't quite get what it is. Sometimes even we don't necessarily get what it is. When people don't get it, and they've got their immediate concerns, it's hard for them to see the relevance. In ILM's case, the immediate concern was to make a movie, and we truly weren't relevant to the job at hand.

Because of that, what we needed was protection at that early stage. That's what we'd had at the University of Utah. ARPA was essentially coming in and protecting "the new," even though we didn't know what "the new" was. It's kind of a hard concept because when most people talk about "the new," they're actually talking about it after the fact. They look back and say how brilliant you were at seeing all this, and so forth. Well, it's all nonsense. When it is new, you don't know it. You're creating something for the future, and you don't know exactly what it is. It's hard to protect that. What we got from George was that protection.

The reason I was thinking of that model was that we had a software project here at Pixar to come up with the next generation of tools and assigned that task to a development group. But we had a different problem. The people responsible for our films didn't look at that development group as this sort of odd thing that somebody else was paying for; they looked at the group as a source of smart people that they could use for the film. We had given the development group a charter to come up with new software, and a year later I found out the whole group had been subverted into providing tools for the existing production.

HANRAHAN: I see. So, they didn't get quite enough protection?

CATMULL: They didn't get enough protection, so I started it up again and put a different person in charge. A year later I came back and found that group had been entirely subverted by production again. So, I thought, "OK, the forces here are much more powerful than I realized."

When we did this the third time we actually put in really strong mechanisms, so basically we set it up to protect it. We also did one other thing: we brought in a person from the outside who was very experienced in delivering bulletproof software. As we went through this, however, we found that everything was on schedule, but that the deliverables were shrinking in order to stay on schedule. At some point you don't deliver enough to make a film, and in that case the schedule slips. We had someone to keep things on schedule, but he didn't want to deliver the software until it was perfect.

Now we had gone to the opposite extreme, where the protection was keeping it isolated from engaging with the user. That's when we put Eben Ostby in charge because Eben knows what it means to engage. Now we're going through the bloody process of taking new software and putting it in production. But we've been through this before, and we know it's a painful, messy process.

To me the trick is that you've got to realize you have two extremes—full engagement and full protection—and you have to know what you're doing to be able to move back and forth between the two. For me, R&D is something you really need to protect, but you don't set it up with an impermeable wall. There comes a time when you need to go into the messy arena, where you actually begin to engage.

HANRAHAN: It seems this is true not just in your business but for any software project.

CATMULL: Yes, this idea can be applied everywhere.

HANRAHAN: Among the many things that are inspiring about Pixar, and one way you've had a huge impact on the world, is that you changed many people's views of what computing is all about. A lot of people think of computing as number crunching whose main application is business and engineering. Pixar added an artistic side to computing. I've talked to many students who realize that art can be part of computing; that creativity can be part of computing; that they can merge their interests in art and science. They think of computing as a very fulfilling pursuit.

ON MOVING TO LUCASFILMS:

What we had—and this is very important—was *protection*. When you're doing something new, it's actually fairly fragile. People don't quite get what it is. Sometimes even we don't necessarily get what it is. When people don't get it, and they've got their immediate concerns, it's hard for them to see the relevance.

I think you've inspired them because you have these incredible artistic people here, and you have incredible technologists here, and you obviously have an interest in both. What's your view on how art and science interact in a place like Pixar?

CATMULL: Two things pop into my mind. The first one comes from being in a position where I can see world-class people on both the art and technical sides. With both groups of people, there's a creative axis and there's an organization axis of managing and making things happen. If you look at these criteria, the distribution of creative and organization skills is the same in both groups. People might think that artists are less organized. It turns out it's all nonsense.

HANRAHAN: I agree completely. Most people think scientists are these really precise, rational, organized people, and artists are these imaginative, emotional, unpredictable people.

CATMULL: What you find is that some artists actually are that way, and some are extremely precise and know what they want. They're organized. They lead others. They're inspirational.

HANRAHAN: There's an incredible craft to it, too. Both the craft of programming and the craft of art can be very detailed and precise.

CATMULL: If you think about the craft of laying out a major software system, you have an architect, and you have a lot of people contributing to it. Well, in a film the director is an architect who is orchestrating contributions from a lot of people and seeing how it all fits together. The organizational skills to do that are similar. We have production people on films. Well, we have production managers in software who help organize and put things together. They're not writing the code, but they're making sure that the people work together and that they're communicating. There are good ones and bad ones, but the structure is the same.

And just as you can have a bug in software, you can also have a bug in

your story. You look and say, "Well, gee, that's stupid!" or "That doesn't make any sense!" Well, yeah, it's a bug!

My second observation about the interaction of art and science is related to the early days of Disney when filmmaking and animation were brand new. This was also part of a technical revolution at that time. They had to figure out how to do color and sound and matting and so forth. They were working out all of those things for many years before the technology matured. Now people look back historically and all they see is the art that came out of it. Very few people pay attention to the role of the changing technology and the excitement of what went on there.

When computer graphics became practical, it reintroduced technical change into this field and invigorated it. I believe that technical change is an important part of keeping this industry vital and healthy.

Yet the tendency of most people is to try to get to a stable place. They just want the right process, which I think is the wrong goal. You actually want to be in a place where you are continually changing things. We're writing our new software system now—strictly speaking, we don't have to do that. I believe the primary reason for doing it is to change what we're doing. We're keeping ourselves off balance, and that's difficult to explain to people. Most people don't want to be in an unstable place; they want to go to the comfort zone, so you're fighting the natural inclinations of most people when you say where we want to be is a place that's unstable.

HANRAHAN: You're in a very unusual situation to have so much art and science mixed together. It would be nice if software companies and technology companies had more of those two kinds of people involved.

At Stanford we have an arts initiative in place right now, and one reason it's popular is not because everybody is going to become an artist, but because everybody should learn about art and the processes that artists use, such as drawing and sketching and brainstorming. We think that even if you're a mechanical engineer or a computer scientist, if you get exposed to the arts, you'll be more innovative. Art adds so much to technology and science just by encouraging a few different ways of thinking.

CATMULL: Here are the things I would say in support of that. One of them, which I think is really important—and this is true especially of the elementary schools—is that training in drawing is teaching people to observe.

HANRAHAN: Which is what you want in scientists, right?

CATMULL: That's right. Or doctors or lawyers. You want people who are observant. I think most people were not trained under artists, so they have an incorrect image of what an artist actually does. There's a complete disconnect with what they do. But there are places where this understanding comes across, such as in that famous book by Betty Edwards [*Drawing on the Right Side of the Brain*].

The notion is that if you learn to draw, it doesn't necessarily mean that you are an artist. In fact, you can learn how to play basketball, but that doesn't mean you can play for the Lakers. But there's a skill you can learn. The same is true with art. This is a skill to learn—for observation, for communication—that we all should have.

The second thing is that there is a notion whereby creativity is applied only to the arts. Those of us who are in the technical areas realize the creative component in those areas. The things that make people creative are very much the same in both fields. We may have different underlying skills or backgrounds, but the notion of letting go and opening up to the new applies in the same way.

HANRAHAN: When I used to run into you, you would always be carrying around some math book, and I always got the sense you wanted to work on a few more cool math and technical problems. Did you ever have a chance to do that? I know you're busy, but is there anything like that you're looking forward to working on?

CATMULL: Well, yes, there still is one. I spent a lot of time trying to get an intuitive feeling for complex numbers. Even though I got good grades in complex analysis when I was in college, I always felt that I was missing some feeling for it. I wanted to have the feeling for how it worked, but all the books explained things in the same way. Part of it was the terminology; for example, I think the word *imaginary* is the wrong word. If they called them *rotation numbers* it would have been an easier thing to get. Basically, I was trying to get at it, and I couldn't find it in a book, although I did find one exception: *Visual Complex Analysis*, by Tristan Needham [Oxford University Press, 1999]. He had been a student of Roger Penrose [at the Mathematical Institute].

HANRAHAN: I remember you pointed that book out. That is a fabulous book.

CATMULL: The unfortunate thing was that he got pulled into administration. I felt he should have written another book, because he was such an inspiration in the way he thought about those things.

HANRAHAN: It was such a visual book, too. It had lots of great diagrams. You got the essence of the ideas; it wasn't just a bunch of derivations.

CATMULL: I got concepts from it that I never got from any other work on analysis. So, I sat down and wrote a program to try to understand this space. I graphed complex series, and then I would look at the complex plane. In doing that, I began to see things, and understand things, that I didn't see before.

I recognize that a lot of physicists actually get that. It took a while to realize that some of the terminology was just about the rotation, but because of the way they approached it, I didn't make that mental leap to get there.

Then I was trying to get to the next place, which was: What does it mean to think about matrices of complex numbers? That's when I ran out of time, but there were some interesting things that I wanted to do in hyperbolic spaces having to do with relativity and physics. For years I wanted to do that, and I still believe there's something there. C

Related articles on queue.acm.org

A Conversation with Kurt Akeley and Pat Hanrahan
http://queue.acm.org/detail.cfm?id=1365496

A Conversation with David Shaw
http://queue.acm.org/detail.cfm?id=1614441

Future Graphics Architectures
William Mark
http://queue.acm.org/detail.cfm?id=1365501

DOI:10.1145/1859204.1859222

Article development led by acmqueue
queue.acm.org

These days, cybercriminals are looking to steal more than just banking information.

BY MACHE CREEGER

The Theft of Business Innovation

An ACM-BCS Roundtable on Threats to Global Competitiveness

VALUABLE INFORMATION ASSETS stretch more broadly than just bank accounts, financial-services transactions, or secret, patentable inventions. In many cases, everything that defines a successful business model (email, spreadsheets, word-processing documents, among others) resides on one or more directly or indirectly Internet-connected personal computers resides in corporate databases, in software that implements business practices, or collectively on thousands of TCP/IP-enabled real-time plant controllers. While not the traditional high-powered information repositories one normally thinks of as attractive intellectual property targets, these systems do represent a complete knowledge set of a business' operations. Criminals, who have come to understand that these information assets have very real value, have set up mechanisms to steal and resell them, bringing great financial harm to their original owners.

In this new world, businesses that may have taken five to six years of trial and error to develop a profitable model are targeted by bad actors who drain and distill operational knowledge from sources not traditionally viewed as highly important. They then resell it to a global competitor who, without having to invest the equivalent time and money, can set up shop and reap its benefits from day one.

In this CTO Roundtable, our joint ACM and BCS-The Chartered Institute for IT panel of security and policy experts discuss how the current threat environment has evolved and the implications for loss in this new environment. At stake is nothing less than the compromise of detailed operational blueprints of the value-creation process. The implications reach far beyond individual businesses, potentially to entire industries and overall economies.

—Mache Creeger

Participants

Louise Bennett, chair, BCS Security Strategic Panel

David J. Bianco, incident handler, General Electric Computer Incident Response Team (GE-CIRT)

Scott Borg, director, chief economist, CEO, U.S. Cyber Consequences Unit

Andy Clark, head of forensics, Detica; Fellow, BCS; Fellow of the Institution of Engineering and Technology

Jeremy Epstein, senior computer scientist, Computer Science Laboratory, SRI International

Jim Norton, visiting professor of electronic engineering, Sheffield University; vice president and Fellow, BCS; chair, BCS Professionalism Board

Steve Bourne, CTO, El Dorado Ventures; past president, ACM; chair, *ACM Queue* Editorial Board; chair, ACM Professions Board

Mache Creeger (moderator) principal, Emergent Technology Associates

CREEGER: While past definitions have narrowly defined valued information as banking codes or secret inventions, criminals have broadened that definition to where they can clone entire businesses through the comprehensive theft of more mundane information such as manufacturing processes, suppliers, customers, factory layout, contract terms, general know-how, and so on. This new shift has significant implications to the competitive balance of entire industries, regardless of company size, and it has implications across the global economic landscape. How do you see this new security threat evolving, and how should businesses respond?

Roundtable panel from bottom left: Jim Norton, David Bianco, Mache Creeger, Louise Bennett; top left: Steve Bourne, Scott Borg, Andy Clark, Jeremy Epstein, and BCS Director for Professionalism Adam Thilthorpe.

PHOTOGRAPHS BY MARJAN SADOUGHI

BORG: In 2004, when the U.S. Cyber Consequences Unit started, we were concerned about intrusions into critical infrastructure facilities, such as chemical plants and refineries. We believed that some of these intrusions were reconnaissance in preparation for an attack that would cause physical destruction.

We got that one wrong, because there were no major attacks of that kind. Further investigation indicated that when the attackers were in the control networks, they gave equal attention to equipment regardless of its ability to blow things up.

What they were doing was copying every bit of operational plant data they could get their hands on: how everything was connected, all the control systems, and settings for every pressure and temperature switch and valve across the entire facility. They were not stealing traditional intellectual property such as trade secrets or proprietary processes but the plant's entire operational workflow.

Soon after these attacks, new facilities in those very industries were popping up in Southeast Asia. No visitors were allowed, and we believe it's because they were exact replicas of attacked facilities.

From an economic standpoint, the degree to which you are ahead of your competition determines how much money you're going to capture from the market. The value reaped from being ahead is very dependent on your lead time as you develop your manufacturing facility. As a rule of thumb, when you open a new facility, you can reduce costs by 5% to 15% each year of operation for roughly the first six years. This amounts to a huge drop in cost, and for a lot of industries represents the majority of the profits. If someone can steal all the operational information it took you six years to develop and open a facility that on day one has the exact same level of efficiency, they have effectively stolen the majority of the profit for your facility.

What is being stolen is something enormously more valuable than what has been lost to credit card or bank fraud. This is a huge issue and puts these companies and potentially entire domestic industries in jeopardy of survival.

NORTON: Should we assume that the attackers also lift one or two key staff people to help interpret this information?

BORG: If you take Asia as an example, using this type of information is often limited by the availability of people who understand Western business practices. This is not something you learn by taking a course locally. To use the information effectively, you have to send someone not only to study in the West but also to work in Western industry.

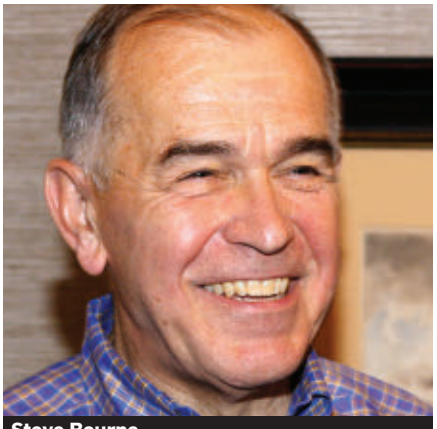
BIANCO: It used to be that you had to be just secure enough that an attacker would give up and go to a less-secure competitor. This is no longer true. Being targeted today means you have something of specific value, and the attackers will probably not go away until they get it. This is fundamentally different from past practices. The people who learned about this in January when Google made its Gmail announcement are probably several years behind everyone else.

CLARK: Much of the business community looks at security as being the people who make sure all the doors and windows are locked. Rarely are security processes aligned with the business, but it's the business that drives security, and security should protect and support valued business processes. That's easier said than done.

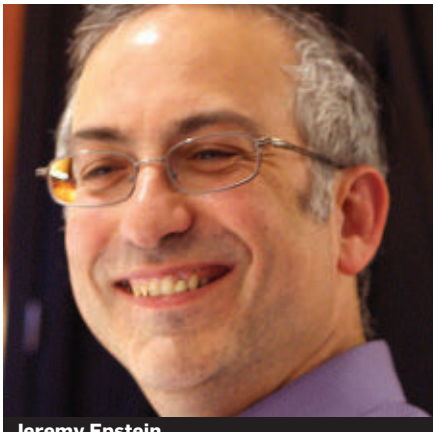
There is also the ethical dilemma of assuming that my competitor and I do business in the same way. That is clearly asymmetric, because your competitor may not follow your business rules. It's hard enough to run a business, be ethical, and work within your regulatory framework without an actor coming in outside of that framework.

We need to: (a) educate people that


Mache Creeger


Steve Bourne


Louise Bennett


Jeremy Epstein

there are others who work outside their business and ethical framework; and (b) define security as a function that works for you by supporting value-creating business processes.

BENNETT: One of the key considerations is motivation. The two main business-attack motivations are money/greed and reputation. People behave differently according to their motivation and the type of business they're attacking. Stealing factory operations know-how is different from stealing information about the pricing of a product that's about to be launched. Both of these are very different from destroying the competition by destroying its reputation.

As a business owner, you have to ask, "In what ways am I vulnerable in the electronic world? Who could attack me, why would they want to, and what would they want to do?"

CLARK: Is it the feeling around the table that industry is more sensitized to the confidentiality associated with cyber attack, rather than treating availability and integrity as equally important issues?

BORG: Companies are sensitive to the confidentiality of the information they designate as intellectual property. They are not as sensitive to the confidentiality of their control systems, their corporate email messages, or just about anything else they are doing. They do not appreciate the scale of the loss they can suffer from that other information being accessed by an outsider.

Jim Lewis (http://csis.org/expert/james-andrew-lewis) tells about a relatively small regional furniture company—not a business you think of as having key intellectual properties—that became an international target. This company had its furniture designs stolen by a Southeast Asian furniture manufacturer that went on to undercut the price.

If you look at your company from an attacker's viewpoint, then you can usually tell whether your company is a target and what specifically would be attractive. It is all about market-sector leadership, anyplace where the company stands out—for example, technology, cost, style and fashion, or even aggressive market expansion.

CLARK: Many of our adversaries play a very long game and do it very well. In the U.S./U.K. style of business we get caught up in quarterly or annual metrics and are not well educated in the long game. Are we naïve in not thinking more about the long game?

BORG: We have many people representing companies who are not properly incentivized to work in the company's long-term best interest. They are compensated on how they did that quarter or that year and not on whether their actions will cause serious crisis four or five years down the road.

CLARK: What advice can we give to IT managers and business leaders to mitigate these threats? Part of the answer is that we need broader education about the nature of threats and we need to understand the long game. I see the attacks on the furniture manufacturer as a long-game play. One waits until the target is ripe for picking, takes it, and moves on.

CREEGER: There is a lot of inertia to making changes in IT to address these issues. What suggestions do you have that would empower an IT person to say to management: "The survival and success of our business depends on you listening to my issues and acting on my recommendations."

NORTON: It has to be done through examples, and people don't want to publicize their attack problems.

BORG: My organization has been warned that we can tell these stories, but if we ever get specific enough that someone can identify one of these companies, then the executives of those companies will be sued by the shareholders, the executives will sue us, the supposed beneficiaries of the attacks will sue, and their business partners will sue as well.

CREEGER: How are we going share our collective wisdom?

NORTON: We can use the airline industry as a model. If you're a pilot of a plane that has a near miss, you can create an anonymous statement about what happened, when it happened, and so on.

CLARK: I can give an example of a breached business that was responsible for placing contractors in high-tech organizations. All of its data was based around individual CVs. An employee at that company chose to extract that data using a USB flash

PHOTOGRAPHS BY MARJAN SADOUGHI

drive and form a new business. The attacker waited until a contractor's previous engagement was coming up for renewal and then remarketed to the contractor under the new business banner.

In this attack, the techniques were very simple. To protect your business, you need to keep a close view of the people in the organization, their motivation and interests, make sure they are satisfied in what they do, and minimize the risk of them doing something criminal and damaging. The focus needs to move from being almost completely technological to a balance of social and technical.

BORG: A common delusion among high-tech companies is that their information has a limited shelf life because they are generating so much new information all the time. This leads to the conclusion that it doesn't matter if people steal information because it's almost immediately obsolete. From an economic standpoint this is just wrong.

NORTON: Let's presume that we can never keep these people out. How do we deal with that?

CLARK: This whole issue of information theft really isn't very new. These issues have been in play for hundreds of years. In our time, some things have changed, most notably pace and data volume stolen. The time necessary to undertake a successful attack has been reduced, and the volume of data that can be taken has dramatically increased.

CREEGER: Can we learn from other fields and experiences? On a previous security panel people talked about addressing the risk of malware infection along the same lines as public health. They said that malware attacks are like the flu. You are never going eradicate it and must live with some ongoing percentage of infections. It will be a different flu every year, and you can minimize your infection risk by implementing certain hygiene protocols.

CLARK: Many people design systems on the assumption they will always work perfectly. Often, auditing features are minimal, sometimes added as a later feature. We need to architect systems on the assumption that breaches will occur, so the functions needed for a proper response are readily available when it happens.

> SCOTT BORG
>
> **There are five steps to follow to carry out a successful cyber attack: find the target; penetrate it; co-opt it; conceal what you have done long enough for it to have an effect; and do something that can't be reversed.**

BIANCO: You should assume all preventative controls will fail. While you still need prevention, you should put your new efforts into detection and response—both in mechanisms and personnel. When prevention fails, if you can't detect failure, you have a very large problem.

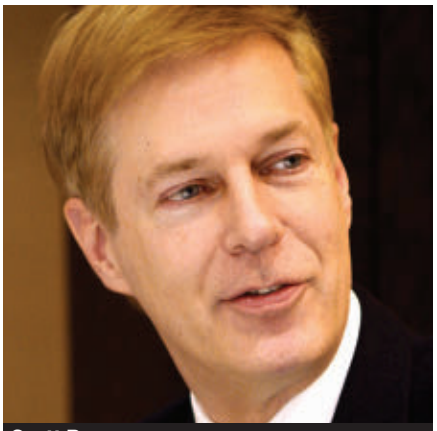
CLARK: In individual applications, we can quickly focus on technical detection without looking at readily available metadata—that includes other systems—that would dramatically improve detection. For example: "Did person A log onto a network? If yes, where was person A when he or she logged on? Does that match what the physical-access-control log reports?"

There is a very real danger that many vendors will provide a good but narrow view of your network and miss the larger context that states, for example, that a user was not supposed to be able to log in from an undetermined physical location.

At Detica we have found real value in mining substantial levels of contextual data that corroborate not just what's happening in the network but what was happening with the individuals that access the network at that point in time. People should not be lulled into a sense of false security because they have purchased a specific niche security product.

CREEGER: Are you saying that we have to start building a huge metadata infrastructure to determine if one event is consistent within a greater context? Who is going to write all these consistency rules that will flag events out of sync with expectations? Who is going to run all these services and on what platforms? How do we architect cost-effective solutions that expend additional cycles to monitor, audit, and determine to the second, third, fourth level whether the person's actually doing what's expected?

BORG: What you are describing as a problem is a huge opportunity. There are five steps you have to follow to carry out a successful cyber attack: find the target; penetrate it; co-opt it; conceal what you have done long enough for it to have an effect; and do something that can't be reversed. Each of these is an opportunity to stop an attacker.


Scott Borg


David J. Bianco


Jim Norton


Andy Clark

You can use these five steps to generate a comprehensive risk chart. By listing all the components of your information system such as hardware, software, networks, and so on, you can itemize the corresponding attack tools and their countermeasures. In this way you can produce a comprehensive security risk grid. Using this methodology to review various sites, we find that while attack tools are spread uniformly across the chart, defensive measures are piled into just a few areas. People typically put almost all their effort into penetration prevention and backup. Most of the other components have no defensive measure to offset defined threats.

We have a huge opportunity for the security industry to develop tools to do such things as quickly identifying bad behavior. Because bad behavior is highly specific to context and industry, security companies need to define industry-specific anomaly detection templates. I believe that is the only way that Andy [Clark's] issues will be resolved.

BENNETT: We have been looking at what I would call the positive side of the attack—that is, the attacker wants to *get* something. There is also an important negative side to an attack in which an attacker is trying to *destroy* something—whether it is reputation, data integrity, or the like. For me, destruction of digital assets is of greater importance.

BORG: The security industry has failed midsize businesses. They don't have the products they need, and they face challenges they can't meet. We have to provide them, either by national policy or security industry initiative, with better solutions than we have right now.

EPSTEIN: One role of government is to react and respond, but the other role is to regulate—to force companies to pay attention to these issues. For all the GEs in the world that are trying to do a good job addressing these issues, there are many more companies that do nothing because nobody is forcing them to do otherwise. In some cases they are independent software vendors selling poorly designed software that can cause future problems.

BORG: This is not an area that can be solved by governmental edict because the technology, including the attack technology, is changing so fast. When the government decides to force people, it has to decide what it's going to force them to do. By the time a standard is identified and regulators have begun enforcement, it not only will be obsolete, but also may very well be an impediment to implementing necessary measures. The best that government can offer is to help the market work better, by making sure there are adequate incentives, adequate product information, and other conditions markets need to function properly.

EPSTEIN: Government also provides the legal infrastructure that allows people to disclaim responsibility. If you buy any commercial security product, the vendor basically says that whatever happens, you are on your own. The ISVs take no responsibility for anything bad that happens.

NORTON: It's tempting to go down that route, but if you are not careful you could destroy the open source community.

CREEGER: Who is accountable? At the end of the day, someone has to stand up and say that this security product meets some reasonably well-understood, generally accepted security standard. For anyone going to a security-focused trade show, it is like a Middle Eastern bazaar of all sorts of competing product schemes.

BENNETT: The responsibility has to lie with the board of the company. One of the big problems, particularly in medium-size companies, is that they have been toddling along for a reasonable length of time and *n* generations of IT have happened while they've been operating. The challenge for the board of that company is to be educated as to what is really required to reduce the threat risk to an acceptable level for their industry.

NORTON: That is why the honey trap is so useful. If you can show that, despite existing protections, a company can still be penetrated, the board ought to be concerned.

CLARK: Many companies with a relatively young board and outlook don't care. They are much more focused on: "We need to do this work now. We are quite high paced and by the time the threat materializes, it's past and I'm not interested." They believe that their

PHOTOGRAPHS BY MARJAN SADOUGHI

business will last for a year or so and then they will move on. We need to be careful of the models we're using and should not claim them to be universal.

BORG: You can identify certain categories, and as you do so, you are also providing the business with clues as to what needs protection. If you are a cost leader, you have to think about what makes you a cost leader and try to secure those things. If you are a technological innovator but not a cost leader, you have a very different focus on what systems you should be trying to protect.

CLARK: Would it be fair to say that while the defense industrial base has been the prime target over the past 10 years, things are clearly changing now?

BORG: One of the real problems with this subject, with this whole field, is it's so hard to keep on top of it. Eighteen months ago, military contractors were overwhelmingly the leading target. That has now shifted to a host of other companies.

We are hearing about companies in South Korea, Indonesia, and other countries that are being offered business research services that will provide them with profiles of competitors and detailed advice on the state of the art in certain growth industries. Many of these research services are selling information they obtain through cyber attacks.

How does this marketplace work? Often there are black-market Web sites that offer the services and have customer reviews and satisfaction ratings.

CREEGER: Are there any concrete examples of industries being cloned?

BORG: Until relatively recently, the main organizations carrying out this kind of activity were national intelligence agencies. They were probably spending millions of dollars to steal the information from one of these target companies. They tried many, many generations of malware, as well as many different attack vectors. We now have privatization of these original efforts—spin-offs from the original national intelligence efforts working for hire.

We are talking about an illegal service—something that's not being sold as a one-off product. We're talking about a sustained business relationship where the customer starts out by buying information for a few thousand dollars (U.S.), becomes gradually convinced of the criminal organization's "integrity," and then goes on to make larger, more strategic purchases.

ANDY CLARK

The structure of available worldwide attack services is broadly commoditized. You can pay using credit cards, not necessarily your own, to buy yourself a worldwide attack service.

My organization has been theorizing about ways to subvert these criminal markets. Just as you can use cyber attacks to undermine trust and damage legitimate markets, you can use those same techniques, including cyber attacks, to undermine criminal markets.

CLARK: The structure of available worldwide attack services is broadly commoditized. You can pay using credit cards, not necessarily your own, to buy yourself a worldwide attack service.

CREEGER: We have learned that business sophistication and marketing in these criminal areas rival anything seen in the legitimate world. As an IT manager, what should I focus on in the next one to three years?

BIANCO: Focus on hiring people who understand how this stuff works.

NORTON: Get people to raise their eyes, look around, and ask, "What is unusual, and how was it caused?"

BORG: In addition, your company needs to be running Symantec, McAfee, Trend Micro, or another retail Internet security package. In many cases, it needs to be hiring the services of an intrusion-detection specialist.

Also, the company has to look at what it is trying to protect: "What are the attackers' motives, what are they going try to break into, what are they after, and what do you need to defend?" Basically, you have to answer the question, "Are you a target, and why?"

CREEGER: You're all strongly saying that the IT people need to be thought of as much more than just the people akin to supporting the plumbing, electrical, and telephone system. IT needs to take a much more integral role in the company's operations and contribute to how the company faces both challenges and opportunities.

NORTON: IT should be engaged in the business and understand how the business works.

CLARK: With regard to the urgency of this issue, I mentioned earlier that the pace of attacks has increased dra-

matically. Because in-place human assets are no longer required, the time needed to penetrate the whole of the industrial base is significantly less than it was in the 1980s. The man running his furniture business might not think he will ever be a target because he is ranked so low, but the attackers will get to him, probably sooner rather than later. We are in danger of thinking about this as a low-paced environment when the reality is it's high paced.

BORG: Cybercrime develops within predictable places. Typical markers are where you have unemployed people with a high level of technical training, where there is an ideological rationale for the attack—because criminals like to feel good about themselves—and where there is some kind of criminal organization to seed the effort.

As these pockets take hold, they often specialize in particular industries or even particular companies. So, a given, very famous, company will tell you that most of its attacks come out of a specific country. There is an opportunity for tampering with the ecology of the attackers and making their lives harder.

CLARK: At a minimum, all businesses should implement a basic level of protection using established commercial products and services. Even though there are many vendors in the market who deliver the basics and do it very well, many companies still do not have basic protection.

Then the next stage is to say, "I've done the basics. Now I need to understand whether I am in this next level and a target."

CREEGER: Given the current mantra of putting things in the cloud, does that make you more secure?

EPSTEIN: Yes and no. I would argue that for small companies and maybe even midsize companies, on balance, it's a good thing. For that sector, it is probably the first time that they're getting some level of professional management and some opportunity for the 24/7 monitoring they clearly need. For large companies, it's probably a huge step backward.

CREEGER: Because it's a one-size-fits-all security model?

NORTON: You have to have some basic quality criteria in the cloud providers.

EPSTEIN: You need to have a way for those small and medium-size companies to discern what type of security those cloud providers provide. A company I worked with outsourced its human-resources system, including all its sensitive employee information, to a cloud provider. I saw the administrator log in using a four-character password, and I said, "You know, this isn't a good idea." An employee, overhearing this, tried to log in with the stock ticker symbol, was successful, and was almost terminated for pointing out the vulnerability. The cloud provider should almost certainly shoulder some of the fault because it turned out that the policy was to accept a minimum of two-character passwords, even for the administrator account. The risk was increased because of the cloud, but the cloud provider was delegating the responsibility to the customer, who didn't have the proper expertise.

CREEGER: What I'm hearing is that the bad guys are way ahead because they're more innovative and profit driven. For the good guys, it's buyer beware, and you must really try to understand your business' realistic vulnerabilities. Always practice basic hygiene and look to the security industry for products such as intrusion protection, firewalls, antivirus, and the like. Don't count on that really bailing you out, however, if you are the target of a sophisticated and determined attacker.

The best advice is to recognize what makes you unique in the market and think honestly about how to protect those assets. This might include spending some money on a computer-literate consultant who could actually help you think through that process.

BORG: You have to guard against having the security consultant sell you a universal solution that promises to secure everything. You need to have a specific strategy that addresses your valued information assets.

CREEGER: Over time, the legitimate computer security world will catch up, and cloud service providers will have tiered certifications designed to fit the needs of specific industry sectors.

CLARK: Yes, but the threat will have moved on. We need to address the fundamental asymmetry of this issue. You will never catch up.

EPSTEIN: I want to add outsourced penetration testing as one more thing to be done. Penetration testing does not tell you where your problems are or how many problems you have but how screwed up you are. Gary McGraw calls it a "badness-ometer." Penetration testing is something that you can take to the board to show real risk and vulnerability.

CLARK: One needs to be cautious and balanced about the way those findings are presented. Penetration testers always find something. It is important that people understand the context of what is found, distinguish what is important in addressing the issues raised, and get to a known baseline. The computer industry should help educate people how their risk profile ranks with similar organizations.

BORG: Employees should never be told to protect valuable assets. If they're told this, they usually protect an object that may be expensive to replace but is not what creates or could destroy value. How value is created is a business' most important asset, and that is what people must focus their protection resources on.

CREEGER: Maybe a recommendation would be to take senior management to an off-site meeting and ask, "If you were a determined attacker to our business, what would you do to damage it or to re-create its value for some other set of shareholders?"

BORG: When we investigate the vulnerabilities of companies, we always get the engineers to sit down and red-team their own company.

CLARK: If you take a slice across the whole company and not just senior management, you'll get much more value. You need an entire cross section of expertise and viewpoints.

NORTON: It's not just about technology but a balance of people, process, and technology. There is intelligence at every level in your organization. The lower levels are often untapped and usually really understand where organizational vulnerabilities reside.

BIANCO: You have to have the right people on staff for this kind of effort. You need to deploy business-specific monitoring technologies and employ someone knowledgeable to look at the output of those systems.

Also, don't be afraid to talk to other

folks in your industry that are being exposed to the same threats. While they may be competitors on the business side, you all have a vested interest in lowering the industrywide threat level. The bad guys talk all the time. If you don't have industry-specific contacts, you will be at an even larger disadvantage. It's probably the least expensive thing you can do to increase your security posture.

BENNETT: Businesses need to understand an attacker's motivation to steal know-how, systems, and other assets. While the typical goal is to replicate and/or destroy the business, the protection of a business' reputation and the rigorous understanding of a business' vulnerabilities are not given the board-level visibility they require. New, young businesses actually understand this better than many medium-size, older businesses.

Attacks may not be just about money and may not be rationally motivated. A motivation for someone to destroy your business may not be "You lose, I win," but "You lose, I stay the same." Given the current state of the recession, that cannot be discounted.

CLARK: When things go wrong, you need to be in a position to understand what happened. That includes not just the technological side but the motivation side as well.

The IT security function needs to have a seat at the management table and directly align with the business' goals. I asked a CISO (chief information security officer) for a large multinational corporation about his objectives. He said, "My first objective is associated with contributing to the financial success of my business." That really focused his mind about the profitability and success of the organization and made him a critical player in the achievement of that goal.

EPSTEIN: Too many organizations spend their information-security resources on protecting their firewalls and other fairly low-level things such as the protocol stack. The activity these days is all happening in the application layer. While a lot of the small and medium-size organizations are just now getting around to protecting the bottom layer, the bottom isn't where the problems are anymore.

If you look at the nature of network attacks, Microsoft, Cisco, etc. have done a reasonably good job. Just because they have pushed attackers higher in the service stack, however, doesn't mean the game is over for us defenders. We have to move our defenses higher as well. We can't just monitor firewall logs anymore. We now have to monitor application logs, and a lot of applications don't have logs. While boards have been hearing the mantra of antivirus, firewall, etc., they now need to understand that the threat has moved up the stack, and the defenses have to move there as well.

I think the cloud is, on the whole, a positive thing. As computer scientists, we need to come up with a way to give users advice on how to select a cloud provider. We need the equivalent of *Consumer Reports* for cloud providers supporting specific industries, especially for small and medium-size businesses.

CREEGER: My take-away is that security is really tied up intimately with the semantics of your business. For a long time, most people have treated security with a one-size-fits solution, usually putting fences around certain critical components without thinking about the real semantics of operations. My impressions from our conversation is not only do IT people need a real seat at the senior management table so they can make substantive contributions to its profitability, but they also need to understand the company's long-term strategy and operations intimately in order to avoid calamity.

MACHE CREEGER

The best advice is to recognize what makes you unique in the market and think honestly about how to protect those assets. This might include spending some money on a computer-literate consultant who could actually help you think through that process.

Related articles on queue.acm.org

Lessons from the Letter
Kode Vicious
http://queue.acm.org/detail.cfm?id=1837255

Intellectual Property and Software Piracy: an interview with Aladdin vice president Gregg Gronowski
http://queue.acm.org/detail.cfm?id=1388781

CTO Roundtable: Malware Defense
http://queue.acm.org/detail.cfm?id=1731902

Mache Creeger (mache@creeger.com) is a technology industry veteran based in Silicon Valley. Along with being a columnist for *ACM Queue*, he is the principal of Emergent Technology Associates, marketing and business development consultants to technology companies worldwide.

DOI:10.1145/1859204.1859226

Only if the programmer can prove (through formal machine-checkable proofs) it is free of bugs with respect to a claim of dependability.

BY ZHONG SHAO

Certified Software

COMPUTER SOFTWARE IS one of the most influential technologies ever created. Software has entered every aspect of our lives, used to control everything from computing and communication devices (such as computers, networks, cellphones, and Web browsers), to consumer products (such as cameras, TVs, and refrigerators), to cyber-physical systems (such as automobiles, medical devices, and aviation systems), and to critical infrastructure (such as financial, energy, communications, transportation, and national defense).

Unfortunately, software is also sometimes our least dependable engineering artifact. Software companies lack the kind of meaningful warranty most other engineering organizations are expected to provide. Major corporations and government agencies worldwide invest in fixing software bugs, but prospects for building reliable software are bleak. The pervasive presence of software bugs also makes all existing computing and information systems vulnerable to security and privacy attacks.

An important cause of such difficulty is the sheer complexity of the software itself. If each line of code is viewed as an individual component, software systems are easily the most complicated things we humans have ever built. Unlike hardware components, software execution can easily lead to an unbounded number of states, so testing and model-checking techniques cannot guarantee reliability. As the hardware community moves deep into new multi-core and cyber-physical platforms, and as software is thoroughly integrated into everyday objects and activities, the complexity of future software could get much worse, even as demand for dependable software becomes more urgent.

For most of today's software, especially low-level forms like operating systems, nobody knows precisely when, how, and why they actually work. They lack rigorous formal specifications and were developed mostly by large teams of developers using programming languages and libraries with imprecise semantics. Even if the original developers had a good informal understanding of the inner work-

» key insights

- **The dependability of a software system should be treated separately from its execution environment; the former is a rigorous mathematical entity, but the latter is imperfect and far less rigorous.**
- **Building end-to-end certified software requires a rich metalogic for expressiveness, a set of domain-specific program logics for modularity and automation, a certified linking framework for interoperability, and machine-checkable proofs for scalability.**
- **The trusted computing base of a good certified framework should contain only components whose soundness and integrity can also be validated by independent third parties.**

ings, their knowledge and assumptions about system behavior (often implicit) are easily lost or broken in subsequent development or maintenance phases.

The software research community has sought to tackle these problems in recent years but remains hampered by three key difficulties:

Lack of metrics. Metrics are still lacking for measuring software dependability, making it difficult to compare different techniques and build steady progress in the field. Dependability often includes attributes like reliability, safety, availability, and security. A system's availability can be measured retroactively as a percentage of its uptime in a given year; for example, 99.9999% means 31.5 seconds downtime per year, but quantifying other attributes is much more difficult. A program with one bug is not necessarily 10 times more secure than a program with 10 bugs. A system's reliability depends on its formal specification, which is often nonexistent.

Worse, software dependability is often confused with the dependability of the software's execution environment, which consists of not just hardware devices but also human operators and the physical world. Since the dependability of the execution environment is often beyond human control, many people view software as a complex biological system, rather than as a rigorous mathematical entity;

System software. A software application's dependability also relies on the dependability of its underlying system software, including OS kernel, device driver, hypervisor, garbage collector, and compiler. These low-level programs are often profoundly complex and bug-prone, but little has been done to make them truly dependable. For example, if an OS kernel or even a compiler has security holes, the entire system could be compromised, regardless of what software developers do at a higher level[19,31]; and

Last-mile problem. Despite recent progress in formal-methods research,

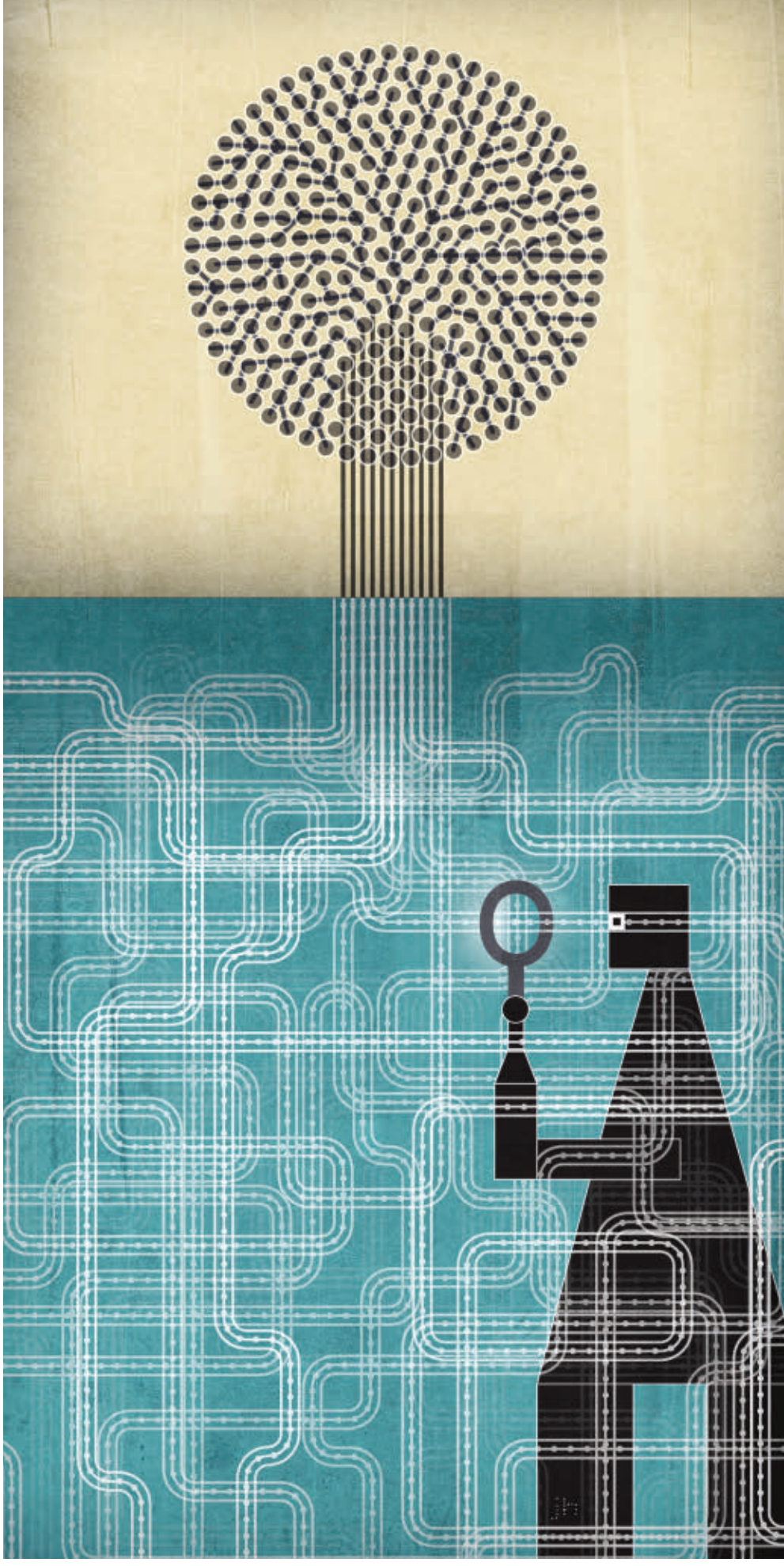


ILLUSTRATION BY JOHN HERSEY

program verification still involves a vexing "last-mile problem." Most software-verification research concentrates on high-level models rather than on actual programs—valuable for finding bugs but leaving a big gap that must be closed before meaningful dependability claims can be made about actual software. Failure to reason about actual code also has serious implications for maintainability; for example, it is difficult for programmers to pinpoint the source and a fix when a new bug is identified and ensure that subsequent updates (to actual code) will not break the code's high-level model.

Leading research on certified software aims to tackle all three. For example, concerning the lack of good metrics, a line is drawn between the actual machine-executable software and the surrounding physical environment (such as hardware devices and human operators). We can neither predict the future of the physical world nor formally certify human behavior, but at least under a well-defined, stable hardware platform (such as the x86 instruction set), the behavior of each machine executable is a rigorous mathematical entity. With a formal specification stating its desirable behavior, we can (at least in theory) rigorously "certify" that the machine executable behaves as expected. A good dependability metric is then just the formal claim developers make and certify about each program.

The long-term goal for research on certified software is to turn code—often a system's weakest link—into its most dependable component. The formal specification given may not precisely capture the behavior of the physical environment, so the overall system may still not function properly, but at least when a problem occurs, programmers and users alike are assured that the behavior of the software is properly documented and rigorously enforced. The specifications for functional correctness of individual components may occasionally be too large to be comprehensible, but many systemwide safety, liveness, and security properties can be stated succinctly and certified with full confidence.

To address the second and third difficulties, software developers must also certify the actual system-software code. Most needed is a new "certified" computing platform where programmers have firm control over the behavior of its system software stack, including bootloader, OS kernel, device driver, hypervisor, and other runtime services. Software consisting of mostly certified components would be easier to maintain, because the effects of updating a certified component would be easier to track, and new bugs would quickly be localized down to the noncertified modules.

Software dependability is often confused with the dependability of the software's execution environment, which consists of not just hardware devices but also human operators and the physical world.

Constructing large-scale certified software systems is itself a challenge. Still unknown is whether it can be done at all and whether it can be a practical technology for building truly dependable software. In this article, I explore this new field, describing several exciting recent advances and challenging open problems.

What It Is

Certified software consists of a machine-executable program C plus a rigorous formal proof P (checkable by computer) that the software is free of bugs with respect to a particular dependability claim S. Both the proof P and the specification S are written using a general-purpose mathematical logic, the same logic ordinary programmers use in reasoning every day. The logic is also a programming language; everything written in logic, including proofs and specifications, can be developed using software tools (such as proof assistants, automated theorem provers, and certifying compilers). Proofs can be checked automatically for correctness—on a computer—by a small program called a proof checker. As long as the logic used by programmers is consistent, and the dependability specification describes what end users want, programmers can be sure the underlying software is free of bugs with respect to the specification.

The work on certified software fits well into the Verified Software Initiative (VSI) proposed by Hoare and Misra[14] but differs in several distinct ways from traditional program-verification systems:

First, certified software stresses use of an expressive general-purpose metalogic and explicit machine-checkable proofs to support modular reasoning and scale program verification to han-

dle all kinds of low-level code.[3,10,24,32] Using a rich mechanized metalogic allows programmers to define new customized "domain-specific" logics (together with its meta theory), apply them to certify different software components, and link everything to build end-to-end certified software.[7] With machine-checkable proofs, proof-checking is automated and requires no outside assumptions. As long as the metalogic is consistent, the validity of proof *P* immediately establishes that the behavior of program *C* satisfies specification *S*.

Existing verification systems often use a rather restricted assertion language (such as first-order logic) to facilitate automation but do not provide explicit machine-checkable proof objects. Program components verified using different program logics or type systems cannot be linked together to make meaningful end-to-end dependability claims about the whole software system. These problems make it more difficult for independent third parties to validate claims of dependability.

Second, with an expressive metalogic, certified software can be used to establish all kinds of dependability claims, from simple type-safety properties to more advanced safety, liveness, security, and correctness properties. Building these proofs need not follow Hoare-style reasoning[15]; much of the earlier work on proof-carrying code[23] constructed safety proofs automatically using such technologies as type-preserving compilation[29,30] and typed assembly language.[22] However, most traditional program verifiers concentrate on partial correctness properties only.

Third, certified software emphasizes proving properties for the actual machine executables, rather than their high-level counterparts, though proofs can still be constructed at the high level, then propagated down to the machine-code level using a certifying or certified compiler. On the other hand, most existing program verifiers target high-level source programs.

Fourth, to establish a rigorous dependability metric, certified software aims to minimize the trusted computing base, or TCB—the small part of a verification framework in which any error can subvert a claim of end-to-end dependability. TCB is a well-known concept in verification and security, as well as a source of confusion and controversy.[5]

The dependability of a computing system rests on the dependable behavior of its underlying hardware devices, human operators, and software. Many program verifiers are comfortable with placing complex software artifacts (such as theorem provers, OS, and compilers) in the TCB because it seems that the TCB of any verification system must include those "hard-to-reason-about" components (such as hardware devices and human operators) so is already quite large.

All program-verification systems are able to create a formal model about the underlying execution environment. Any theorem proved regarding the software is with respect to the formal model only, so the TCB for any claim made regarding the software alone should not include hardware devices and human operators.

Still, any bug in the TCB would (by definition) compromise the credibility of the underlying verification system. A smaller TCB is generally more desirable, but size is not necessarily the best indicator; for example, a 200-line garbage collector is not necessarily more reliable than a 2,000-line straightforward pretty printer. The TCB of a good certified framework must include only components whose soundness and integrity can also be validated by independent third parties.

Components of a certified framework. A typical certified framework (see Figure 1) consists of five components:

The certified software itself. Including both machine code and formal proof;

Formal machine model. Providing the operational semantics for all machine instructions;

Formal dependability claim for the software. Including safety property, security policy, and functional specification for correctness;

Underlying mechanized metalogic (not shown). For coding all proofs, specifications, and machine-level programs; and

Proof checker. For checking the validity of all the proofs following the inference rules of the metalogic.

If the proof of a given certified software package can be validated by the proof checker, then execution of the software on the formal machine model is guaranteed to satisfy a formal dependability claim.

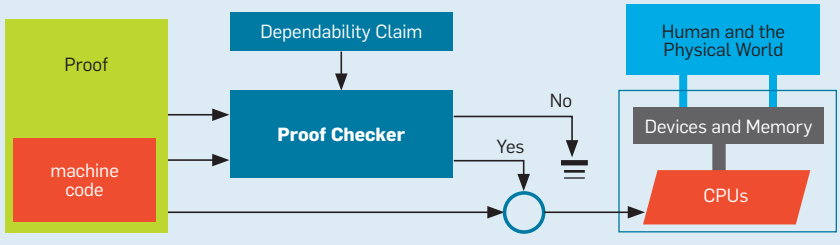
Things can still, however, go wrong. First, the mechanized metalogic could be inconsistent, a risk that can be minimized if the framework designers choose a simple, well-understood, general-purpose metalogic and prove (perhaps on paper) why it is indeed consistent.

Second, the proof checker is a computer program, so it could go wrong all by itself. But if the framework uses a simple logic with a small number of inference rules, the proof checker can be made quite small, written in assembly, and verified by hand.

Third, the formal machine model might not reflect hardware behavior. Most hardware vendors perform intensive hardware verification, so this risk can be minimized if hardware and software developers share the machine specifications. Even if not possible, the framework designer can still validate the model by comparing its operational semantics with the instruction-set reference manuals.

Finally, the formal dependability specification (*SP*) may not accurately

Figure 1. Components of a certified framework.



capture the behavior of the human or physical world. Nevertheless, *SP* is formally stated and the code is guaranteed to satisfy *SP*. Here, I deliberately decoupled the correctness of verification from the specification process. Existing efforts validating and testing specifications are, of course, valuable and complementary to the certification process.

Since a dependability claim is made regarding only the formal machine model, the TCB of such a certified framework consists of just the consistency proof of the metalogic and the integrity of the proof checker, both of which should be demonstrable by independent third parties (such as through the peer-review process of a top-quality journal). If the computer science community would agree on a single metalogic (a good thing), this task of standardizing a metalogic would need to be done only once. Certified software would then no longer be the weakest link in a dependable system.

Mechanized metalogic. A key enabling technology for certified software is to write formal proofs and specifications as typed functional programs, then have a computer automatically check the validity of the proofs, in the same way a static type-checker does type-checking. This idea came from the well-known Curry-Howard correspondence referring to the generalization of a syntactic analogy between systems of formal logic and computational calculi first discovered by the American logicians Haskell Curry and William Howard. Most advances for developing large-scale machine-checkable proofs were made only during the past 10 years; see an excellent survey by Barendregt and Geuvers[2] and a 2008 overview article by Hales.[11]

In the context of certified software, a few more requirements must be addressed: The logic must be consistent and expressive so software developers can express everything they want to say. It must also support explicit machine-checkable proof objects and be simple enough that the proof checker can be hand-verified for correctness.

Because software components may be developed using different programming languages and certified using different domain-specific logics and type systems, mechanized metalogic must also support meta-reasoning. It can be used to represent the syntax, inference rules, and meta-proofs (for their soundness) of the specialized object logics.

Much of the current work on certified software is carried out in the Coq proof assistant.[16] Coq itself provides a rich higher-order logic with powerful inductive definitions, both crucial to writing modular proofs and expressive specifications.

Advantages. With certified software, the dependability of a software system would be measured by the actual formal dependability claim it is able to certify. Because the claim comes with a formal proof, the dependability can be checked independently and automatically in an extremely reliable way.

A formal dependability claim can range from making almost no guarantee, to simple type-safety property, to deep liveness, security, and to correctness properties. It provides a great metric for comparing different techniques and making steady progress toward the system's overall dependability.

If the software community would agree on a metalogic and work out the formal models of a few popular computing platforms, certified software would provide an excellent framework for accumulating dependable software components. Since proofs are incontrovertible mathematical truths, once a software component is certified, its trustworthiness (with respect to its specification) would presumably last for eternity.

Unlike higher-level programming languages, certified software places no restrictions on the efficiency of its underlying code and the way programs are developed. Because the metalogic is as rich as the one programmers use in daily reasoning, and everything running on a computer must eventually be executed as a machine executable, if programmers believe (informally) that their super-efficient and sophisticated code really works as they claim, there should be a way to formally write down their proofs. When dependability is not an issue, the software can be used as is, assuming proper isolation from the rest of the system; when programmers really care about dependability, they must provide the formal machine-checkable proof.

On the other hand, certified software encourages the usual best practices in software engineering and program verification. Certifying large-scale systems clearly benefits from high-level programming abstraction, domain-specific logics, modular decomposition and refinement, model-driven design and development, the correctness-by-construction methodology,[12] and automated theorem-proving tools. The only difference is they now insist on receiving hard evidence (such as machine-checkable proof objects) as a way to deliver quality assurance and measure the effectiveness of the technologies.

Certified software also decouples the proof-construction and program-development tools from the proof-checking infrastructure. The rich metalogic provides the ultimate framework for building up layers of abstraction for complex software. Once they are formed, programmers can build different software components and their proofs using completely different methods. Because specifications and proofs are both represented as programs (within a computer), they can be debugged, updated, transformed, analyzed, and reused by novel proof-engineering tools.

Certified software also significantly improves the maintainability of the underlying system. A local change to an individual component can be checked quickly against its specification, with its effect on the overall system known immediately. A major reorganization of the system can be done in a principled way by comparing the changes against high-level specifications programmers have for each certified component.

Challenges. The main challenge of certified software is the potentially huge cost in constructing its specifications and proofs, though it can be cut dramatically in the following ways:

First, how software is developed makes a big difference in the system's future dependability. If the software is full of bugs or developed without consideration of the desirable dependability claim, post-hoc verification would be extremely expensive in terms of time and money or simply impos-

sible. A proactive approach (such as correctness-by-construction[12]) should lower the cost significantly.

Second, building certified software does not mean that programmers must verify the correctness of every component or algorithm used in its code; for example, in micro-kernels or virtual-machine monitors, it is often possible for programmers to verify a small set of components that in turn perform run-time enforcement of security properties on other components.[33]

Dynamic validation (such as translation validation for compiler correctness[26]) also simplifies proofs significantly; for example, it may be extremely difficult to verify that a sophisticated algorithm A always takes an input X and generates an output Y such that $R(X, Y)$ holds; instead, a programmer could extend A by adding an additional validation phase, or validator, that checks whether the input X and the output Y indeed satisfy the predicate R, assuming R is decidable. If this check fails, the programmer can invoke an easier-to-verify (though probably less-efficient) version of the algorithm A. To build certified software, all the programmer needs to do is certify the correctness of the validator and the easier version of the algorithm, with no need to verify algorithm A anymore.

Third, the very idea that proofs and specifications can be represented as programs (within a computer) means that developers should be able to exploit the synergy between engineering proofs and writing large programs, building a large number of tools and proof infrastructures to make proof construction much easier.

Finally, formal proofs for certified software ought to be much simpler and less sophisticated than those used in formal mathematics.[11] Software developers often use rather elementary proof methods to carry out informal reasoning of their code. Proofs for software are more tedious but also more amenable for automatic generation.[6,28]

Certified software also involves other challenges. For example, time to market is likely terrible, assuming dependability is not a concern, so the cost of certification would be justified only if end users truly value a dependability guarantee. Deployment would be difficult since most real-world engineers do not know how to write formal specifications, let alone proofs. Pervasive certification requires fundamental changes to every phase in most existing software-development practices, something few organizations are able to undertake. The success of certified software critically relies on efforts initially developed in the research community.

Since proofs are incontrovertible mathematical truths, once a software component is certified, its trustworthiness (with respect to its specification) would presumably last for eternity.

Recent Advances

Advances over the past few years in certified software have been powered by advances in programming languages, compilers, formal semantics, proof assistants, and program verification. Here, I sample a few of these efforts and describe the remaining challenges for delivering certified software:

Proof-carrying code. Necula's and Lee's 1996 work[23] on proof-carrying code (PCC) is the immediate precursor to the large body of more recent work on certified software. PCC made a compelling case for the importance of having explicit witness, or formal machine-checkable evidence, in such applications as secure mobile code and safe OS kernel extensions. PCC allows a code producer to provide a (compiled) program to a host, along with a formal proof of safety. The host specifies a safety policy and a set of axioms for reasoning about safety; the producer's proof must be in terms of these axioms.

PCC relies on the same formal methods as program verification but has the advantage that proving safety properties is much easier than program correctness. The producer's formal proof does not, in general, prove the code produces a correct or meaningful result but does guarantee execution of the code satisfies the desirable safety policy.

Checking proofs is an automated process about as simple as programming-language type-checking; on the other hand, finding proofs of theorems is, in general, intractable. Subsequent work on PCC focused on building a realistic certifying compiler[4] that automatically constructs proofs (for simple type-safety properties) for a large subset of Java and on reducing the size of proof witness, an important

concern in the context of mobile code.

An important PCC advantage inherited by certified software is that the software does not require a particular compiler. As long as the code producer provides the proof, the code consumer is assured of safety. This significantly increases the flexibility available to system designers.

The PCC framework is itself quite general, but the original PCC systems suffered from several major limitations: Most notable was that the proof checker had to rely on a rather specific set of typing rules so did not support more expressive program properties; the typing rules were also error-prone, with their soundness often not proved, so a single bug could undermine the integrity of the entire PCC system.

Foundational PCC, or FPCC,[1,13] tackled these problems by constructing and verifying its proofs using a metalogic, with no type-specific axioms. However, FPCC concentrated on building semantic models for high-level type-safe languages, rather than performing general program verification.

Certified assembly programming. CAP[32] is a logic-based approach for carrying out general program verification inside a rich mechanized metalogic (such as the one provided by Coq). Like Hoare logic, a CAP program consists of assembly code annotated with pre- and post-conditions and program invariants. Unlike traditional Hoare-style verification, all CAP language constructs (such as assembly instruction sets), program assertions, inference rules, operational semantics, and soundness proofs are implemented inside the mechanized metalogic. This design makes it possible to build a complete certified software package with formal dependability-claim and machine-checkable proofs. With help from a proof assistant, programmers are able to combine manually developed proof scripts with automated proof tactics and theorem provers, allowing CAP to support verification of even undecidable program properties.

CAP marries type-based FPCC with Hoare-style program verification, leading to great synergy in terms of modularity and expressiveness. Hoare logic is well known for its limited support for higher-order features; most Hoare systems do not even support verification of simple type-safety properties. However, both shortcomings are easily overcome in type-based approaches. Subsequent work on CAP over the past five years developed new specialized program logics for reasoning about such low-level constructs as embedded code pointers,[24] stack-based control abstractions,[10] self-modifying code,[3] and garbage collectors.[21]

Machine-checkable proofs are necessary for allowing third parties to quickly establish that a software system indeed satisfies a desirable dependability claim.

Under type-based FPCC, function returns and exception handlers are often treated as first-class functions, as in continuation-passing style (CPS), even though they have more limited scope than general first-class continuations. For functional programmers, CPS-based code is conceptually simple but requires complex higher-order reasoning of explicit code pointers (and closures). For example, if a function needs to jump to a return address (treated as continuation), the function must assert that the return address is indeed a valid code pointer to jump to. But the function does not know exactly what the return address will be, so it must abstract over properties of all possible return addresses, something difficult to do in first-order logic.

In our work on stack-based control abstraction,[10] my colleagues and I showed that return addresses (or exception handlers) are much more disciplined than general first-class code pointers; a return address is always associated with some logical control stack, the validity of which can be established statically; a function can cut to any return address if it establishes the validity of its associated logical control stack. Such safe cutting to any return address allows programmers to certify the implementation of sophisticated stack operations (such as `setjmp/longjmp`, weak continuations, general stack cutting, and context switches) without resorting to CPS-based reasoning. For example, when programmers certify the body of a function, they do not need to treat its return address as a code pointer; all they need is to make sure that at the return, the control is transferred to the original return address. It is the caller's responsibility to set up a safe return address or valid code pointer; this is much easier because a caller

often knows the return address that must be used.

Local reasoning and separation logic. Modular reasoning is the key technique for making program verification scale. Development of a certified software system would benefit from a top-down approach where programmers first work out the high-level design and specification, then decompose the entire system into smaller modules, refine high-level specifications into actual implementation, and finally certify each component and link everything together into a complete system.

However, there is yet another critical dimension to making program verification modular. Traditional Hoare logics often use program specifications with arbitrarily large "footprints." Separation logic[17,27] advocates "local reasoning" using small-footprint specifications; that is, the specification of each module (or procedure) should refer only to data structures actually touched by a module's underlying code. By concisely specifying the separation of heap and other resources, separation logic provides succinct yet powerful inference rules for reasoning about shared mutable data structures and pointer anti-aliasing.

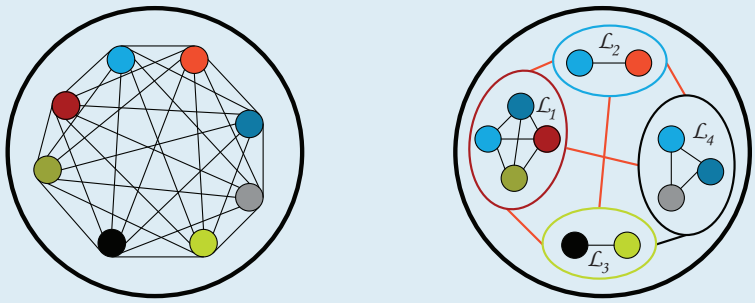
Concurrent separation logic (CSL)[25] applies the same idea to reasoning about shared-memory concurrent programs, assuming the invariant that there always exists a partition of memory among different concurrent entities and that each entity can access only its own part of memory. This assumption might seem simple but is surprisingly powerful. There are two important points about the invariant: First, the partition is logical; programmers do not need to change their model of the physical machine, which has only one global shared data heap, and the logical partition can be enforced through separation logic primitives. Second, the partition is not static and can be adjusted dynamically during program execution by transferring the ownership of memory from one entity to the other.

Under CSL, a shared-memory program can be certified as if it were a sequential program since it is always manipulating its private heap; to access shared memory, it must invoke an atomic operation that transfers resources between the shared heap and the local heap. Several recent efforts have extended CSL with rely-guarantee reasoning, so even lock-free concurrent code can be certified using modular small-footprint specifications.

Domain-specific logics and certified linking. A key first step toward making certified software practical is to show it is possible to carry out end-to-end certification of a complete software system. Large software systems, especially low-level system software, use many different language features and span many different abstraction levels. For example, the Yale FLINT group's (http://flint.cs.yale.edu) ongoing project[8] to verify a simplified OS kernel exposes such challenges. In it, the kernel includes a simple bootloader, kernel-level threads and a thread scheduler, synchronization primitives, hardware interrupt handlers, and a simplified keyboard driver. Although it has only 1,300 lines of x86 assembly code, it uses dynamic code loading, thread scheduling, context switching, concurrency, hardware interrupts, device drivers, and I/O. How would a programmer use machine-checkable proofs to verify the safety or correctness properties of such a system?

Verifying the whole system in a single program-logic or type system is impractical because, as in Figure 2a, such a verification system would have to consider all possible interactions among these features, including dynamic code loading, concurrency, hardware interrupts, thread scheduling, context switching, and embedded code pointers, many at different abstraction levels. The resulting logic, if it exists, would be highly complex and difficult to use. Fortunately, software developers seem to never use all features simultaneously. Instead, they use only a limited combination of features—at a particular abstraction level—in individual program modules. It

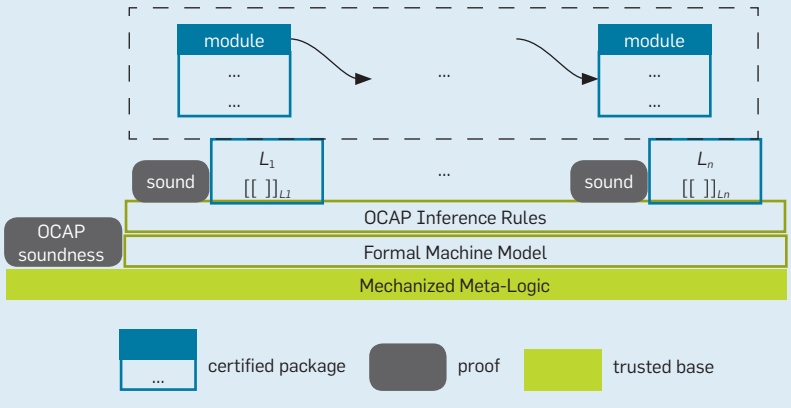
Figure 2. Using domain-specific logics to verify modules.



(a) One-for-All Logic

(b) Domain-Specific Logics

Figure 3. An open framework for building certified software.



would be much simpler to design and use specialized "domain-specific" logics (DSL) to verify individual program modules, as in Figure 2b. For example, for the simplified OS kernel, dynamic code loading is used only in the OS boot loader, and interrupts are always turned off during context switching; embedded code pointers are not needed if context switching can be implemented as a stack-based control abstraction.

To allow interactions of modules and build a complete certified software system, programmers must also support interoperability of different logics. In 2007, my colleagues and I developed a new open framework for CAP, or OCAP,[9] to support verification using specialized program logics and for certified linking of low-level heterogeneous components. OCAP lays a set of Hoare-style inference rules over the raw operational semantics of a machine language (see Figure 3), and the soundness of these rules is proved in a mechanized metalogic so it is not in the TCB. OCAP uses an extensible and heterogeneous program-specification language based on the higher-order logic provided by Coq. OCAP rules are expressive enough to embed most existing verification systems for low-level code. OCAP assertions can be used to specify invariants enforced in most type systems and program logics (such as memory safety, well-formedness of stacks, and noninterference between concurrent threads). The soundness of OCAP ensures these invariants are maintained when foreign systems are embedded in the framework.

To embed a specialized verification system $\mathcal{L}$, OCAP developers must first define an interpretation $[\![\ ]\!]_{\mathcal{L}}$ that maps specifications in $\mathcal{L}$ into OCAP assertions; they then prove system-specific rules/axioms as lemmas based on the interpretation and OCAP rules. Proofs constructed in each system can be incorporated as OCAP proofs and linked to compose the complete proof.

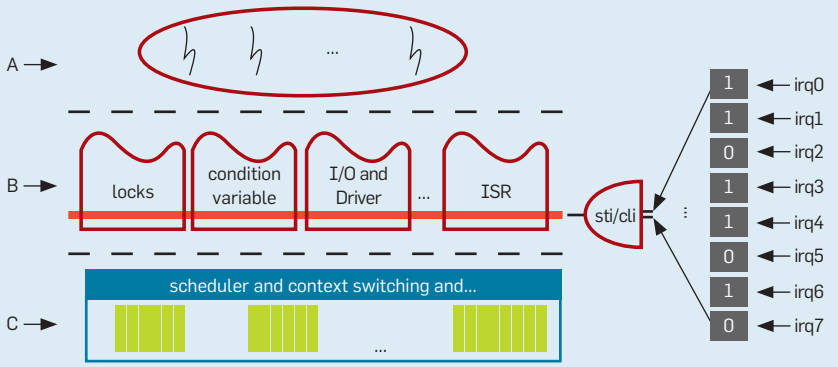
There are still many open issues concerning OCAP design: For example, to reason about information-flow properties, it must provide a semantic-preserving interpretation of high-order types (in an operational setting). And to support liveness properties, it must support temporal reasoning of program traces.

Certified garbage collectors and thread libraries. In 2007, my colleagues and I used OCAP to certify several applications involving both user-program code and low-level runtime code. In one application,[9] we successfully linked programs in typed assembly language (TAL)[22] with a certified memory-management library. TAL supports only type-preserving memory updates; the free memory is invisible to TAL code. We certified the memory-management library in stack-based CAP, or SCAP,[10] supporting reasoning about operations over free memory while ensuring that the invariants of TAL code are maintained.

Also in 2007, in another application,[21] we developed a general framework for certifying a range of garbage collectors and their mutators. If we had tried to develop a single type system to type-check both an ML-like type-safe language and the underlying garbage collector (requiring fancy runtime type analysis), the result would have involved analyzing polymorphic types, which is extremely complex. However, the ML type system never needs to know about runtime tagging and the internals of the garbage collector. Moreover, implementation of the collector need not understand the polymorphic type system used in type-checking ML code; it needs to only distinguish pointers from non-pointers. A better approach, which we followed in 2007, is to certify these modules using different domain-specific logics, thus avoiding the difficult task of designing a universal program logic. Certified garbage collectors can then be linked with certified mutator code to form a complete system.

A year later, in a third application,[8] we successfully certified the partial correctness of a preemptive thread library extracted from our simplified OS kernel. The kernel was implemented in 16-bit x86 assembly and worked in real mode for uniprocessor only. It consisted of thread context switching, scheduling, synchronizations, and hardware interrupt handlers. We stratified the thread implementation by introducing different abstraction layers with well-defined interfaces. In Figure 4, at the highest level (Level A), preemptive threads follow the standard concurrent programming model. The execution of a thread can interleave with or be preempted by other threads. Synchronization operations are treated as primitives. Hardware interrupts are abstracted away and handled at Level B where code involves both hardware interrupts and threads; synchronization primitives, input/output operations, device drivers, and interrupt handlers are all implemented at this level, and interrupt handling is enabled/disabled explicitly using `sti/cli`. At the lowest level (Level C), the thread scheduler and the context-switching routine manipulate the threads' execution contexts stored in thread queues (on the heap). Interrupts are invisible at this level because they are always disabled. Libraries implemented at a lower level are exposed as abstract primitives for the level above it, and their operational semantics in the high-level abstract machine

Figure 4. Decomposition of a preemptive thread implementation.



serve as formal specifications for the low-level implementation.

The stratified system model gives programmers a systematic and principled approach for controlling complexity. Programmers can thus focus on a subset of language features at each level and certify different software components using specialized program logics.

Certified and certifying compilation. Much work in the program-verification community concentrates on source-level programs written in high-level languages (such as C, Java, and C#). In order to turn these programs into certified assembly components suitable for linking in the OCAP framework, OCAP developers must show that their corresponding compiler is also trustworthy.

CompCert is a certified compiler for a subset of C (called C minor, or Cm) developed in 2006 by Leroy.[20] By "certified" compiler, I mean the compiler itself is proved correct. Indeed, Leroy specified formal operational semantics for Cm, as well as for the machine language, building a machine-checkable proof in Coq whereby the compiler preserves behavior from one operational semantics to another. However, the current CompCert compiler supports only sequential Cm programs. It also must be bootstrapped by the OCaml compiler, even though the OCaml compiler is not verified.

On the other hand, a certifying compiler is not necessarily correct but will take a (certified) source program and generate certified assembly code. Much work on certifying compilation focuses on type-safe source languages and can preserve only type-safety properties. A challenging open problem is to extend certifying compilation to preserve deep correctness and security properties.

Lightweight formal methods. Building large-scale certified software systems does not always require heavyweight program verification. Most software systems are built from modular components at several levels of abstraction. At the lowest levels are the kernel and runtime-system components discussed earlier. At the highest levels are components with restricted structure operating on well-defined interfaces. The restricted structure can use a type-safe, high-level programming language with high-level concurrency primitives or C programs (even concurrent C programs) in a style understandable to static-analysis tools. Both restricted styles are in widespread commercial use today.

Lightweight formal methods (such as high-level type systems, specialized program logic, with decidable decision procedure, and static analysis) can help guarantee important safety properties with moderate programmer effort; error messages from the typechecker, decision procedure, and static-analyzer usually give appropriate feedback in the programming process. These properties are sometimes also security properties, as in this example: "Module A cannot read the private variables of module B, except through the public methods provided by B." Using information-flow type systems or static analysis a programmer can obtain a stronger version of the same guarantee while also adding "... and not only that, but the public methods of module B do not leak the value of private variable *x*."

Lightweight formal methods can be used to dramatically cut the cost of building certified software. For a programmer, the challenge is to make them generate explicit proof witness (automatically) and link them to certified low-level kernel and runtime components. With proper embedding, lightweight formal methods would fit nicely into the DSL-centric OCAP framework for constructing end-to-end certified software.

Automation and proof engineering. The end goal of certified software is a machine-checkable dependability metric for high-assurance software systems. Certified software advocates the use of an expressive metalogic to capture deep invariants and support modular verification of arbitrary machine-code components. Machine-checkable proofs are necessary for allowing third parties to quickly establish that a software system indeed satisfies a desirable dependability claim. Automated proof construction is extremely important and desirable but should be done only without violating the overall integrity and expressiveness of the underlying verification system.

Much previous research on verification reflected full automation as a dominating concern and was reasonable if the primary goal is finding bugs and having an immediate effect on the real world's vast quantity of running software. Unfortunately, insisting on full automation also severely hinders the power and applicability of formal verification; many interesting program properties (that end users care about) are often undecidable (full automation is impossible), so human intervention is unavoidable. Low-level program modules often have subtle requirements and invariants that can be specified only through high-order logic; programming libraries verified through first-order specifications often have to be adapted and verified again at different call sites.

Nevertheless, there is still great synergy in combining these two lines of software-verification work. The OCAP framework described earlier emphasizes domain-specific (including decidable first-order) logics to certify the components in a software system. Successful integration would allow programmers to get the best of both lines.

Developing large-scale mechanized proofs and human-readable formal specifications will be an exciting research field on its own, with many open issues. Existing automated theorem provers and Satisfiability Modulo Theories solvers[6] work on only first-order logic, but this limited functionality conflicts with the rich metalogic (often making heavy use of quantifiers) required for modular verification of low-level software. Proof tactics in existing proof assistants (such as Coq) must be written in a different "untyped" language, making it painful to develop large-scale proofs.

Conclusion

Certified software aligns well with a 2007 study on software for dependable systems[18] by the National Research Council (http://sites.nationalacademies.org/NRC/index.htm) that argued for a direct approach to establishing dependability, whereby software developers make explicit the dependability claim and provide direct evidence that the software indeed satisfies the claim. However, the study did not explain

what would make a clear and explicit dependability claim, what would serve as valid evidence, and how to check the underlying software to ensure it really satisfies the claim without suffering credibility problems.[5]

The study also said that the dependability of a computer system relies not only on the dependability of its software but also on the behavior of all other components in the system, including human operators and the surrounding physical environment. Certified software alone cannot guarantee the dependability of the computer system. However, many advantages, as explained earlier, follow from separating the dependability argument for the software from the argument for the software's execution environment.

Computer software is a rigorous mathematical entity for which programmers can formally certify claims of dependability. However, the behavior of human operators depends on too many factors outside mathematics; even if they try hard, they would probably never achieve the kind of rigor they can for software. By focusing on software alone and insisting that all certified software come with explicit machine-checkable proofs, a formal claim of dependability can be used as a metric for measuring software dependability. Formal specifications are also more complete and less ambiguous than informal specifications written in natural languages; this should help human operators better understand the behavior of the underlying software.

A key challenge in building dependable systems is to identify the right requirements and properties for verification and decide how they would contribute to the system's overall dependability. Certified software does not make this task easier. Research on certifying low-level system software would give software developers more insight into how different programming-abstraction layers would work together. Insisting on machine-checkable proof objects would lead to new high-level certified programming tools, modular verification methodologies, and tools for debugging specifications, all of which would make developing dependable software more economical and painless.

Acknowledgments

I would like to thank Xinyu Feng, Daniel Jackson, George Necula, Ashish Agarwal, Ersoy Bayramoglu, Ryan Wisnesky, and the anonymous reviewers for their valuable feedback. C

References

1. Appel, A.W. Foundational proof-carrying code. In *Proceedings of the 16th Annual IEEE Symposium on Logic in Computer Science* (Boston, June 16–19). IEEE Press, Los Alamitos, CA, 2001, 247–258.
2. Barendregt, H.P. and Geuvers, H. Proof-assistants using dependent type systems. In *Handbook of Automated Reasoning*, A. Robinson and A. Voronkov, Eds. Elsevier Scientific Publishing BV, Amsterdam, The Netherlands, 2001, 1149–1238.
3. Cai, H., Shao, S., and Vaynberg, A. Certified self-modifying code. In *Proceedings of the 2007 ACM Conference on Programming Language Design and Implementation* (San Diego, June 10–13). ACM Press, New York, 2007, 66–77.
4. Colby, C., Lee, P., Necula, G., Blau, F., Plesko, M., and Cline, K. A certifying compiler for Java. In *Proceedings of the 2000 ACM Conference on Programming Language Design and Implementation* (Vancouver, B.C., June 18-21). ACM press, New York, 2000, 95–107.
5. DeMillo, R.A., Lipton, R.J., and Perlis, A.J. Social processes and proofs of theorems and programs. In *Proceedings of the Fourth Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages* (Los Angeles, Jan.17–19). ACM Press, New York, 1977, 206–214.
6. de Moura, L.M. and Bjørner, N. Z3: An efficient SMT solver. In *Proceedings of the 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (Vol. 4963 of LNCS)* (Budapest, Mar. 29–Apr. 6). Springer-Verlag, Berlin, 2008, 337–340.
7. Feng, X., Shao, Z., Guo, Y., and Dong, Y. Combining domain-specific and foundational logics to verify complete software systems. In *Proceedings of the Second IFIP Working Conference on Verified Software: Theories, Tools, and Experiments (Vol. 5295 of LNCS)* (Toronto, Oct. 6–9). Springer-Verlag, Berlin, 2008, 54–69.
8. Feng, X., Shao, Z., Dong, Y., and Guo, Y. Certifying low-level programs with hardware interrupts and preemptive threads. In *Proceedings of the 2008 ACM Conference on Programming Language Design and Implementation* (Tucson, AZ, June 10–13). ACM Press, New York, 2008, 170–182.
9. Feng, X., Ni, Z., Shao, Z., and Guo, Y. An open framework for foundational proof carrying code. In *Proceedings of the 2007 ACM SIGPLAN International Workshop on Types in Language Design and Implementation* (Nice, France, Jan. 16). ACM Press, New York, 2007, 67–78.
10. Feng, X., Shao, Z., Vaynberg, A., Xiang, S., and Ni, Z. Modular verification of assembly code with stack-based control abstractions. In *Proceedings of the 2006 ACM Conference on Programming Language Design and Implementation* (Ottawa, June 11–14). ACM Press, New York, 2006, 401–414.
11. Hales, T.C. Formal proof. *Notices of the AMS 55*, 11 (Dec. 2008), 1370–1380.
12. Hall, A. and Chapman, R. Correctness by construction: Developing a commercial secure system. *IEEE Software 19*, 1 (Jan./Feb. 2002), 18–25.
13. Hamid, N.A., Shao, Z., Trifonov, V., Monnier, S., and Ni, Z. A syntactic approach to foundational proof-carrying code. In *Proceedings of the 17th Annual IEEE Symposium on Logic in Computer Science* (Copenhagen, July 22–25). IEEE Press, Los Alamitos, CA 2002, 89–100.
14. Hoare, C.A.R. and Misra, J. Verified software: Theories, tools, experiments. In *Proceedings of the First IFIP Working Conference on Verified Software: Theories, Tools, and Experiments (Vol. 4171 of LNCS)* (Zurich, Oct. 10–13). Springer-Verlag, Berlin 2005, 1–18.
15. Hoare, C.A.R. An axiomatic basis for computer programming. *Commun. ACM 12*, 10 (Oct. 1969), 576–580.
16. Huet, G., Paulin-Mohring, C., et al. *The Coq Proof Assistant Reference Manual. The Coq Release v6.3.1*, May 2000; http://coq.inria.fr
17. Ishtiaq, S. and O'Hearn, P.W. BI as an assertion language for mutable data structures. In *Proceedings of the 28th ACM Symposium on Principles of Programming Languages* (London, Jan. 17–19). ACM Press, New York, 2001, 14–26.
18. Jackson, D., Thomas, M., and Millett, L. *Software for Dependable Systems: Sufficient Evidence?* The National Academies Press, Washington, D.C., 2007.
19. King, S.T., Chen, P.M., Wang, Y.-M., Verbowski, C., Wang, H.J., and Lorch, J. Subvirt: Implementing malware with virtual machines. In *Proceedings of the 2006 IEEE Symposium on Security and Privacy* (Oakland, CA, May 21–24). IEEE Press, Los Alamitos, CA, 2006, 314–327.
20. Leroy, X. Formal certification of a compiler back-end or: Programming a compiler with a proof assistant. In *Proceedings of the 33rd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages* (Charleston, SC, Jan. 11–13). ACM Press, New York, 2006, 42–54.
21. McCreight, A., Shao, Z., Lin, C., and Li, L. A general framework for certifying garbage collectors and their mutators. In *Proceedings of the 2007 ACM Conference on Programming Language Design and Implementation* (San Diego, June 10–13). ACM Press, New York, 2007, 468–479.
22. Morrisett, G., Walker, D., Crary, K., and Glew, N. From System F to typed assembly language. In *Proceedings of the 25th ACM Symposium on Principles of Programming Languages* (San Diego, Jan. 19–21). ACM Press, New York, 1998, 85–97.
23. Necula, G. and Lee, P. Safe kernel extensions without run-time checking. In *Proceedings of the Second USENIX Symposium on Operating System Design and Implementation* (Seattle, Oct. 28–31). USENIX Association, Berkeley, CA, 1996, 229–243.
24. Ni, Z. and Shao, Z. Certified assembly programming with embedded code pointers. In *Proceedings of the 33rd Symposium on Principles of Programming Languages* (Charleston, SC, Jan. 11–13). ACM Press, New York, 2006, 320–333.
25. O'Hearn, P.W. Resources, concurrency and local reasoning. In *Proceedings of the 15th International Conference on Concurrency Theory (Vol. 3170 of LNCS)* (London, Aug. 31–Sept. 3). Spinger-Verlag, Berlin, 2004, 49–67.
26. Pnueli, A., Siegel, M., and Singerman, E. Translation validation. In *Proceedings of the Fourth International Conference on Tools and Algorithms for Construction and Analysis of Systems (Vol. 1384 of LNCS)* (Lisbon, Portugal, Mar. 28–Apr. 4). Springer-Verlag, Berlin 1998, 151–166.
27. Reynolds, J.C. Separation logic: A logic for shared mutable data structures. In *Proceedings of the 17th Annual IEEE Symposium on Logic in Computer Science* (Copenhagen, July 22–25). IEEE Press, Los Alamitos, CA 2002, 55–74.
28. Schulte, W., Xia, S., Smans, J., and Piessens, F. A glimpse of a verifying C compiler. In *Proceedings of the C/C++ Verification Workshop* (Oxford, U.K., July 2, 2007).
29. Shao, Z. An overview of the FLINT/ML compiler. In *Proceedings of the ACM SIGPLAN Workshop on Types in Compilation* (Amsterdam, The Netherlands, June 8, 1997).
30. Tarditi, D., Morrisett, G., Cheng, P., Stone, C., Harper, R., and Lee, P. TIL: A type-directed optimizing compiler for ML. In *Proceedings of the 1996 ACM Conference on Programming Language Design and Implementation* (Philadelphia, May 21–24). ACM Press, New York, 1996, 181–192.
31. Thompson, K. Reflections on trusting trust. *Commun. ACM 27*, 8 (Aug. 1984), 761–763.
32. Yu, D., Hamid, N.A., and Shao, Z. Building certified libraries for PCC: Dynamic storage allocation. In *Proceedings of the 2003 European Symposium on Programming (Vol. 2618 of LNCS)* (Warsaw, Apr. 7–11). Springer-Verlag, Berlin, 2003, 363–379.
33. Zeldovich, N. *Securing Untrustworthy Software Using Information Flow Control.* Ph.D. thesis, Department of Computer Science, Stanford University, Oct. 2007; http://www.cs.stanford.edu/histar/

Zhong Shao (zhong.shao@yale.edu) is a professor in the Department of Computer Science at Yale University, New Haven, CT.

DOI:10.1145/1859204.1859225

What do wikis, blogs, podcasts, social networks, virtual worlds, and the rest do for corporate productivity and management?

BY STEPHEN J. ANDRIOLE

Business Impact of Web 2.0 Technologies

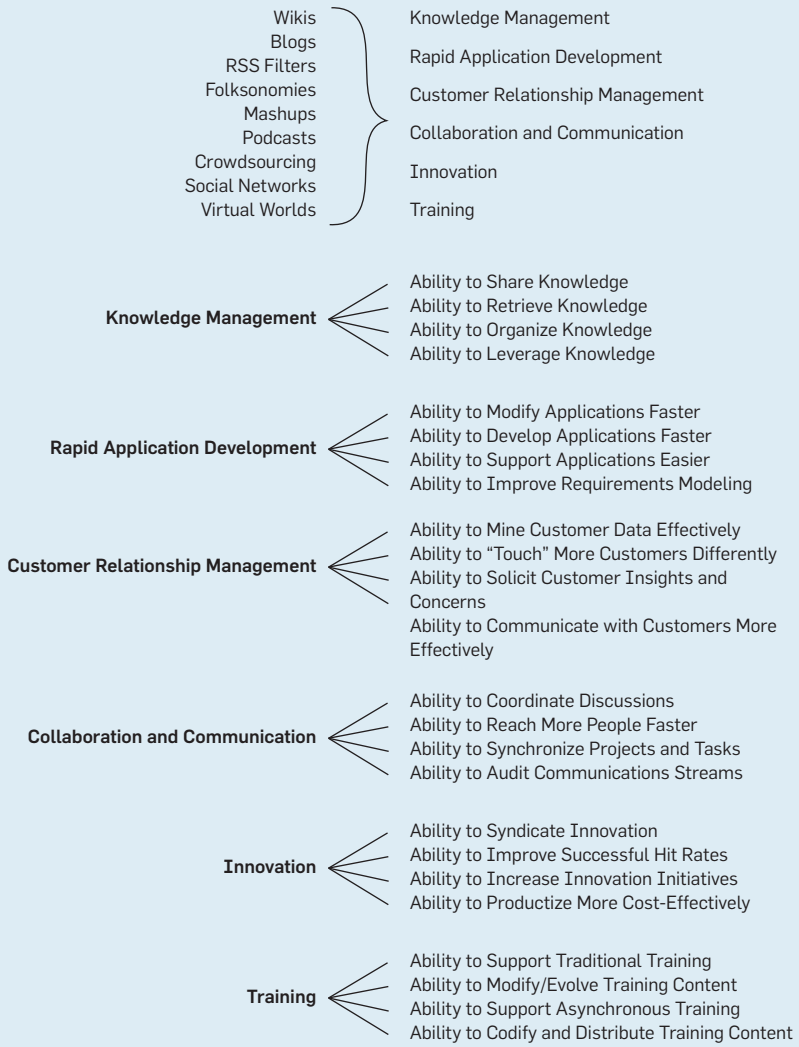
THIS ARTICLE DESCRIBES research designed to measure the impact of the business value of wikis, blogs, podcasts, folksonomies, mashups, social networks, virtual worlds, crowdsourcing, and RSS filters—all Web 2.0 technologies. Properly deployed, they may well permit companies to cost-effectively increase their productivity and, ultimately, their competitive advantage; the research reported here includes results of interview, observation, and survey data-collection from select companies and industries primarily in the U.S. across six performance areas: knowledge management, rapid application development, customer relationship management, collaboration/communication, innovation, and training. The results include caution, skepticism, and a significant contribution to collaboration and communication. Wikis, blogs, and RSS filters have had the greatest impact, while virtual worlds have had virtually none. Security remains a concern, but we found that communication and collaboration are generally well served by Web 2.0 technologies.

Only limited published research is available today exploring the contribution of Web 2.0 technologies to

» key insights

- **Web 2.0 technologies can help improve collaboration and communication within most companies.**
- **These technologies should be assessed to determine real impact, and a number of assessment techniques, including interviews, observations, and surveys, can be used to measure impact over time across multiple business areas.**
- **These technologies can help improve collaboration and communication across multiple vertical industries, though many companies are cautious about deploying them.**

Figure 1. Impact metrics.



corporate productivity and management. Gartner Group (http://www.gartner.com), Forrester Research (http://www.forrester.com), IDC (http://www.idc.com), and the Cutter Consortium (http://www.cutter.com) report that Web 2.0 technologies are rapidly making their way into corporate technology infrastructures and architectures. But the way they are used and the impact they are having have not been reported in a systematic way.

My research posed the following questions to managers and executives:

- What good is Web 2.0 technology to your company?;
- What problems might Web 2.0 technology solve?;
- How can we use the technology to save or make money?; and
- What are the best ways to exploit the technology without complicating existing infrastructures and architectures?

Research objectives included:

- Understand which Web 2.0 tools and techniques are most likely to improve corporate productivity and management;
- Identify how Web 2.0 tools and techniques can be used to enhance corporate productivity and management; and
- Measure impact via collection of interview, direct observational, and survey data.

Questions addressed included:

- Can wikis, blogs, RSS filters, and folksonomies help companies improve their knowledge management?;
- Can wikis be used to build "corporate encyclopedias," training manuals, and other forms of documentation?;
- Can blogs be used to vet ideas about markets, customers, and strategies?;
- Can podcasts be used to document products?;
- Can folksonomies be used to organize structured and unstructured content?;
- Can RSS filters be used to create content streams to improve customer relationship management?;
- Can mashups be used for rapid application development?; and
- Can crowdsourcing be used to stimulate innovation?

Research methods included:

- Profile the range of Web 2.0 technologies available to corporations;
- Define "impact" across multiple dimensions of productivity;
- Collect data on the use of Web 2.0 technologies and the impact areas through interviews, direct observation, and surveys;
- Analyze the data to identify usage patterns and impact;
- Identify correlations from the survey data among technologies and impact areas; and
- Measure the relative impact of individual and groups of technologies on individual and groups of impact areas.

(Figure 1 outlines specific impact metrics.)

Business 2.0, *Fast Company*, *Businessweek*, and other business publications cover Web 2.0 and even Web 3.0, the so-called "new Net" and the next digital gold rush. Is it indeed another bubble, with Web 2.0 (then Web 3.0) vendors crashing and burning like their dot-com predecessors a decade ago? The online trade journal *Web 2.0 Journal* (http://www.web2journal.com) explores all sorts of Web 2.0 technologies, while just about every major technology vendor has released multiple white papers on the promise of Web 2.0 technologies and applications. There are also many Web 2.0 blogs, including Dion Hinchcliffe's Web 2.0 (http://www.web2.socialcomputingmagazine.com), that attract a

growing number of participants. If this were 1999, we'd call Web 2.0 a "killer app" or "disruptive technology." However, we're still not sure today about the business impact of Web 2.0 technologies, which have evolved on the consumer-to-consumer side of the Web. Social networking sites like MySpace (http://www.myspace.com), Facebook (http://www.facebook.com), and Friendster (http://www.friendster.com) were developed to connect individuals anxious to share experiences, photographs, videos, and other personal aspects of their daily lives. These sites grew rapidly with huge amounts of user-created content; YouTube (http://www.youtube.com) is probably the best example of such content.

Our research reflects corporate deployment trends and business impact. Will Web 2.0 technology be widely adopted because it dramatically and cost-effectively improves corporate performance? Will it ultimately disappoint the business and technology professionals it's expected to please?

Interview Questions

The questions we posed to participating companies and that defined our observation included:

- How did you become aware of the availability of Web 2.0 technologies?;
- What is your understanding of how Web 2.0 technologies might positively affect productivity?;
- What is a great Web 2.0 productivity scenario for your company?;
- What's a really bad business scenario for your company trying to exploit Web 2.0 technologies?;
- Which Web 2.0 technologies have your company piloted?;
- Which Web 2.0 technologies have you avoided, and why?;
- What is their impact?;
- How would you quantify the impact of Web 2.0 technologies in the following areas: knowledge management, rapid application development, customer relationship management, collaboration, communication, innovation, and training?;
- What is your company's greatest success with Web 2.0 technologies?;
- What is your company's greatest disappointment?;
- What excites you most about Web 2.0 technologies?;
- What worries you most about investing in these technologies?;
- Which infrastructure or architecture issues worry you most?;
- Does business acceptance worry you?;
- Does IT acceptance worry you?; and
- Where do you think your company will be with Web 2.0 applications in the next three years?

These questions guided our interviews and observation exercises. Our conversations were designed to understand what companies were doing with Web 2.0 technologies, their impact, and their alignment with expectations, fears, and trends. They assumed that companies are in the relatively early stages of their Web 2.0 application deployment, are still learning what the technologies can and cannot do, and are motivated to understand their potential.

Figure 2. Summary interview findings.

	Internally Focused Applications	Externally Focused Applications
Collaboration/Communication	The majority of Web 2.0 technology applications are in this area. Viewed as "safe," they allow companies to pilot them while testing impact on security, infrastructure, total cost of ownership, and intellectual property.	Early adopters pilot Web 2.0 technologies outside the corporate firewall to establish alternative communication and collaboration patterns with employees, suppliers, clients, and customers, permitting improved communication.
Knowledge Management	KM is a natural result of deployment of wikis, blogs, podcasts, and RSS filters. Formal KM tools are giving way to more informal Web 2.0 tools, a trend expected to continue.	KM will support externally focused organizations (such as those in the consulting and retail industries) before internally focused organizations formally adopt it, slowed by concerns over security, privacy, and intellectual property.
Rapid Application Development	Mashup and related technology is gradually replacing more traditional RAD technology. As more and more components, application programming interfaces, and widgets are published, more RAD progress will be made.	RAD tools and techniques will formalize for technology vendors and technology-driven companies and industries, as more and more components, applications programming interfaces, and widgets are published by direct publishers and third-party hosts.
Customer Relationship Management	CRM applications are slow to absorb the extensible abilities of Web 2.0 technologies internally and especially externally. It will take time for Web 2.0 technologies to be integrated with and extended from existing CRM technologies.	CRM is a natural partner for Web 2.0 technologies, especially such tools as RSS filters, podcasts, mashups and blogs. There are countless ways to leverage Web 2.0 technologies on behalf of customers and suppliers, but, due to deployment anxiety, such applications will lag.
Training	Companies increasingly use wikis, blogs, podcasts, and RSS filters for training and education. Their ease of use and participatory nature appeal to a growing number of companies. Relatively low cost helps.	Third-party training and education providers will leverage Web 2.0 technologies, integrating them into the already substantial online training and education industry. The tools will then be sold back to customers to improve learning of all kinds.
Innovation	Web 2.0 technologies have little impact on the innovation process. There are spotty innovation applications of crowdsourcing for R&D and selected applications of folksonomies, RSS filters, and mashups, but the area is generally not affected.	Web 2.0 tools, techniques, and especially attitudes will alter the innovation process in many industries by facilitating direct communication and collaboration among creators and buyers of new products and services, thus shortening the innovation life cycle.

Interviews

We undertook a number of interviews and conversations, combined with direct observation, to determine the deployment of Web 2.0 technologies and, more important, the impact they have on corporate productivity. Our conversations occurred in Q1 and Q2 2008 with companies in the pharmaceutical, chemical, real estate/mortgage, information technology, and financial services industries agreeing to in-depth interviews and access to the teams implementing select Web 2.0 technologies. The interviews were conducted with senior technology managers in each company. Approximately 15 senior managers participated in the interviews.

The five companies represented the following vertical industries:

Company A. Big pharmaceutical company;

Company B. Global chemicals company;

Company C. National real estate and mortgage company;

Company D. Global IT company; and

Company E. Large financial services company.

The questions we asked and the responses included:

▸ How did you become aware of the availability of Web 2.0 technologies?;

Big pharmaceutical company: "Reading; conferences, vendors, and IT staff";

Global chemicals company: "Vendors, IT staff, and business partners";

National real estate and mortgage company: "Vendors and IT staff";

Global IT company: "Competitors, industry publications"; and

Large financial services company: "Trade publications, industry organizations."

▸ What is your understanding of the range of Web 2.0 technologies that might positively affect productivity?:

Big pharmaceutical company: "Primarily blogs, wikis, and podcasts";

Global chemicals company: "Blogs, wikis, podcasts, and RSS";

National real estate and mortgage company: "Blogs, wikis, podcasts, and RSS";

Global IT company: "Blogs, wikis, RSS, and virtual reality"; and

Large financial services company: "Blogs, wikis, mashups, and tagging."

There are serious concerns about intellectual property, proprietary information, privacy, security, and control.

▸ What would be a great Web 2.0 productivity scenario for your company?

Big pharmaceutical company: "Very fast, cheap but productive applications";

Global chemicals company: "Easy to deploy with lots of payback";

National real estate and mortgage company: "Fast, cheap to deploy, with major productivity";

Global IT company: "Integrates well with existing technology"; and

Large financial services company: "Transparent but effective."

▸ What would be a really bad scenario for your company?"

Big pharmaceutical company: "Lots of distraction due to the technology";

Global chemicals company: "Expensive, time-consuming deployment that fails";

National real estate and mortgage company: "Loss of control of the technology";

Global IT company: "Exposure of company secrets"; and

Large financial services company: "Everyone playing around with this stuff when they should be working."

▸ Which Web 2.0 technologies have you piloted?

Big pharmaceutical company: "Wikis and blogs";

Global chemicals company: "Wikis and blogs";

National real estate and mortgage company: "Wikis, RSS, and blogs";

Global IT company: "Wikis, blogs, and RSS filters"; and

Large financial services company: "Wikis, blogs, and mashups."

▸ Which Web 2.0 technologies are you avoiding, and why?

Big pharmaceutical company: "Virtual worlds, stupid";

Global chemicals company: "Virtual worlds, no clue how they might help us";

National real-estate and mortgage company: "Virtual worlds and blogs, way too much data to control";

Global IT company: "Blogs and crowdsourcing, way too much proprietary data in them"; and

Large financial services company: "Social networks, way too distracting during work."

▸ What has been the impact of the

technologies?

Big pharmaceutical company: "Too early to tell, way too early";

Global chemicals company: "Suspicious of trade-offs between 'fun' and 'productivity'";

National real-estate and mortgage company: "Who the hell knows?";

Global IT company: "People seem to like them, but I don't know the real impact"; and

Large financial services company: "We are hopeful."

▸ How would you quantify the impact in knowledge management, rapid application development, customer relationship management, collaboration, communication, innovation and training?

Big pharmaceutical company. "Collaboration and communication is where the action is; this is the real impact we're seeing at this point; plus, there's a lot of user acceptance of wikis, blogs, and social networks; we're getting more formal with KM where wikis and blogs are being used to codify information and vet decisions; only doing a little with RAD and mashups, but that will come in time; same with CRM, where we plan to use the tools to better communicate with customers and suppliers; wikis are emerging as training tools; not too much yet with innovation; a little worried about crowdsourcing outside the firewall";

Global chemicals company. "Wikis and blogs have changed the way we communicate: they're easy and fast, and everyone can participate; KM is fast following improved communications and collaboration; the IT team is crazy about mashups; they are able to build applications very quickly for the business, so I'd say RAD has improved; CRM with external customers and suppliers is behind the other applications; we're a little leery of working outside the firewall with these tools; training is a natural; we're using wikis, blogs, and podcasts for training, with good results; still nothing with virtual worlds or crowdsourcing, a little too 'out there' for us";

National real estate and mortgage company. "We're all over these tools for data and content management; RSS filters are used internally and externally, and we tag everything for better search and access; communication and collaboration are obvious beneficiaries of the tool use; CRM is our next application, where RSS and other content will be provided to our customers; virtual worlds are not there for us yet, but we like wikis, blogs, and podcasts for training; they are cheaper and faster than hiring a training company; innovation is happening inside the company with crowdsourcing and blogs";

Global IT company. "Communication and collaboration have improved since we introduced some Web 2.0 tools; consumerization has definitely taken hold here; people, especially the younger ones, are simply extending their personal experience with the tools to the workplace without missing a beat; KM is just sort of happening on its own, repositories are being built without a formal project to do so; CRM is still not on our radar, though we're doing a lot of things internally we could provide our customers and suppliers; mashup technology is the fastest RAD technology we've ever seen; we're training with wikis and blogs, and the time savings are large"; and

Large financial services company. "Impact has been spotty; I separate fun from productivity; sure, everyone likes these tools, but I'm not convinced that the benefit is there yet;

Table 1. Web 2.0 technology deployment.

Which Web 2.0 technologies have you deployed? (Please select all that apply.)	Response Percent	Response Total
Wikis	62.2%	61
Internal employee blogs	48.0%	47
External customer blogs	20.4%	20
RSS filters	32.7%	32
Folksonomies/ content management	21.4%	21
Mashups	11.2%	11
Virtual worlds	1.0%	1
Internal crowdsourcing	6.1%	6
External crowdsourcing	4.1%	4
Internal social networks	25.5%	25
External social networks	17.3%	17
None	22.4%	22
Other (please specify):	5.1%	5

Table 2. Overall expectations.

How would you rate your expectations about the contribution that Web 2.0 technologies would make to productivity and management?	Response Percent	Response Total
High	23.7%	18
Medium	55.3%	42
Low	21.1%	16
None at all	0%	0

wikis and blogs help communication, especially collaboration, but I wonder just how much; we have so much to do, and even though Web 2.0 tools are pretty easy to use, they still require time and effort; we already have KM tools and databases that permit us to organize and search; we have CRM tools we've invested a ton of money in; we have contractors, vendors, and partners that assist our innovation efforts; and what about the negative impact on security?; we like the CRM aspects of the technologies, but I need to see empirical cost-benefit data before I declare victory."

▸ What is your greatest success with Web 2.0 technologies?

Big pharmaceutical company: "The ability to record knowledge and experiences in a single format and location";

Global chemicals company: "Internal buzz; everybody likes the new stuff";

National real estate and mortgage company: "Wikis are being used for training";

Global IT company: "Using crowdsourcing internally to solve some tough problems"; and

Large financial services company: "Building some RSS filters to better organize information; also using folksonomies to organize data and content."

Table 3. Expectations by impact area.

To which areas did you believe that Web 2.0 technologies would contribute to most? (Please select all that apply.)	Response Percent	Response Total
Knowledge management	78.9%	60
Rapid application development	22.4%	17
Customer relationship management	44.7%	34
Collaboration and communication	90.8%	69
Innovation	46.1%	35
Training	43.4%	33
Other (please specify):	2.6%	2

Table 4. Actual impact data.

To which areas have Web 2.0 technologies contributed the most? (Please select all that apply.)	Response Percent	Response Total
Knowledge management	53.9%	41
Rapid application development	17.1%	13
Customer relationship management	18.4%	14
Collaboration and communication	81.6%	62
Innovation	21.1%	16
Training	7.9%	6
Other (please specify):	2.6%	2

▸ What has been your company's greatest disappointment?

Big pharmaceutical company: "Seeing a lot of what I consider to be sensitive information in wikis, blogs, and podcasts";

Global chemicals company: "IT's inability to control this stuff";

National real estate and mortgage company: "No feedback on what it's good for";

Global IT company: "Lack of vendor support"; and

Large financial services company: "The caution of IT."

▸ What excites your company most about Web 2.0 technologies?

Big pharmaceutical company: "How easy it is to deploy new, useful technology";

Global chemicals company: "How we can displace more expensive technologies for much cheaper and easier-to-use technologies;

National real estate and mortgage company: "How easy it is to use the new stuff";

Global IT company: "How open it is"; and

Large financial services company: "How it extends existing capabilities."

▸ What worries you the most?

Big pharmaceutical company: "Integration with existing technologies";

Global chemicals company: "Integration with business processes";

National real estate and mortgage company: "Support";

Global IT company: "Intellectual property and privacy, a lot"; and

Large financial services company: "Security, privacy, IP, and all of the proprietary data that fills wikis, blogs, crowdsourced solutions, podcasts, and everything else this technology makes transparent."

▸ What infrastructure or architecture issues worry you?

Big pharmaceutical company: "Security, security, and security";

Global chemicals company: "Support";

National real estate and mortgage company: "Governance. Who owns these tools?";

Global IT company: "Integration and interoperability with our applications"; and

Large financial services company: "Integration with our existing appli-

cations and architectures."

▸ Does business acceptance worry you?

Big pharmaceutical company: "Not at all, as long as it works and doesn't cost too much, they will embrace it";

Global chemicals company: "The business always wants to try new things; it's IT that slows things down";

National real estate and mortgage company: "The business is skeptical about all the new tools IT brings to the table, so they'll be cautious";

Global IT company: "The business wants only low-cost solutions"; and

Large financial services company: "If it's free and powerful, they'll love it."

▸ Does IT acceptance worry you?

Big pharmaceutical company: "Yes, they always find something 'wrong' with the new stuff, always worried about support";

Global chemicals company: "No, they are pushing the stuff";

National real estate and mortgage company: "Cost always worries IT; it's been beaten into them over time; so the technology needs to be cheap to deploy and support";

Global IT company: "They will come around; they don't like how easy it is for employees to just set up blogs and wikis, often end-running them"; and

Large financial services company: "They see the business value, or at least the potential in these tools, so I think we are OK here."

▸ Where do you think you will be with Web 2.0 applications in three years?

Big pharmaceutical company: "Fully accepted and integrated";

Global chemicals company: "There, but you need to ask me about Web 3.0 technologies";

National real estate and mortgage company: "Mainstream by that time we will have figured out what to do with them";

Global IT company: "Well-received and productive"; and

Large financial services company: "Still a little skeptical."

Results. The interviews and direct observations revealed consistent trends among the interview subjects (see Figure 2). We learned that Web 2.0 technologies, in spite of the hype, are entering the enterprise slowly but deliberately. The exception is there are clearly applications not entirely controlled by the enterprise's technology organization. The majority of applications are entering organizations in areas where expectations can be managed, costs are low, and tool integration and interoperability (with existing applications and infrastructures) are manageable. We also learned there are serious concerns about intellectual property, proprietary information, privacy, security, and control.

Technology organizations are both advancing and delaying deployment of Web 2.0 technologies. Some absolutely require that Web 2.0 technologies, like all enterprise technologies, be governed by the same processes governing the acquisition, deployment, and support of all digital technologies. Others are loosening their grip somewhat, primarily because they believe it's virtually impossible to prevent business units and project teams from creating wikis and blogs.

There is also a hierarchy of Web 2.0 tools. All companies we interviewed deployed wikis and blogs, and many deployed RSS filters and podcasts.

Table 5. Knowledge management impact data by ability.

In the area of knowledge management, have Web 2.0 technologies contributed to your organization's ability to...	Not at all	Very little	Somewhat	A great deal	Response Total
Share knowledge	3.9% (3)	10.5% (8)	51.3% (39)	34.2% (26)	**76**
Retrieve knowledge	9.2% (7)	13.2% (10)	55.3% (42)	22.4% (17)	**76**
Organize knowledge	6.6% (5)	22.4% (17)	52.6% (40)	18.4% (14)	**76**
Leverage knowledge for problem-solving	13.2% (10)	31.6% (24)	35.5% (27)	19.7% (15)	**76**

Table 6. Web 2.0 technologies and knowledge management.

In terms of improving knowledge management, which Web 2.0 technologies have contributed the most? (Please select all that apply.)	Response Percent	Response Total
Wikis	69.7%	53
Internal employee blogs	30.3%	23
External customer blogs	10.5%	8
RSS filters	13.2%	10
Folksonomies/ content management	18.4%	14
Mashups	3.9%	3
Virtual worlds	1.3%	1
Internal crowdsourcing	2.6%	2
External crowdsourcing	0%	0
Internal social networks	14.5%	11
External social networks	7.9%	6
We have not seen any improvement in knowledge management.	7.9%	6
Other (please specify):	2.6%	2

Table 7. Rapid application development impact data by ability.

In the area of rapid application development, have Web 2.0 technologies contributed to your organization's ability to...

	Not at all	Very little	Somewhat	A great deal	Response Total
Modify applications faster	39.5% (30)	22.4% (17)	30.3% (23)	7.9% (6)	**76**
Develop applications faster	39.5% (30)	23.7% (18)	20.3% (23)	6.6% (5)	**76**
Support applications better	40.8% (31)	22.4% (17)	25.0% (19)	11.8% (9)	**76**
Improve requirements modeling	39.5% (30)	23.7% (18)	28.9% (22)	7.9% (6)	**76**

Table 8. Web 2.0 technologies and rapid application development.

In terms of improving rapid application development, which Web 2.0 technologies have contributed the most? (Please select all that apply.)	Response Percent	Response Total
Wikis	44.7%	34
Internal employee blogs	14.5%	11
External customer blogs	9.2%	7
RSS filters	6.6%	5
Folksonomies/ content management	5.3%	4
Mashups	6.6%	5
Virtual worlds	1.3%	1
Internal crowdsourcing	7.9%	6
External crowdsourcing	0%	0
Internal social networks	7.9%	6
External social networks	0%	0
We have not seen any improvement in rapid application development.	30.3%	23
Other (please specify):	7.9%	6

Fewer deployed social networks, mashups, and folksonomies, and even fewer invested in crowdsourcing and virtual worlds. Deployment momentum is at work, as it often is when new technologies appear. Momentum breeds momentum, and we can expect wikis, blogs, podcasts, and RSS filters to gain momentum as other Web 2.0 technologies lag. The models for exploiting these early-adopted technologies will thus grow faster, wider, and deeper than optimization models for, say, virtual worlds.

Finally, an important distinction separates internal applications from their external counterparts. We noticed that our companies were much more willing to pilot Web 2.0 technologies inside than outside their firewalls, not because they feared failure or wanted to avoid tipping their hands to competitors, but because of deepening concerns about security and access to corporate private data.

Our interviews provided one level of insight into the adoption and impact of Web 2.0 technologies, but what did the survey data provide?

The Survey

The survey questions focused on background issues, impact expectations, and the impact the technologies have across the six areas. The Cutter Consortium, a research and consulting organization, administered the survey to its stable of CIOs, CTOs, CFOs, CEOs, and COOs representing more than 20 vertical industries, including small offices/home offices, small and mid-size businesses, and large global enterprises. The five companies we interviewed also participated in the survey. In addition to these five companies, 93 companies from around the world also responded to the survey.

Results. Table 1 outlines the survey results, along with the deployment landscape. Wikis and blogs lead the charge, followed by RSS filters.[a] Perhaps surprising is the deployment of internal social networks and folksonomies/content management applications. No one seems to like living in a virtual world. The use of external customer blogs is also interesting and suggestive of our desire to reach out to customers any way we can. We must also acknowledge that 22% in the survey did not deploy any Web 2.0 technologies at all.

These results are consistent with our interview data. The most obvious Web 2.0 technologies, including wikis and blogs, are being deployed more rapidly than virtual worlds, crowdsourcing, and mashups. There's caution around early adoption of any new technology. Due to the freewheeling nature of Web 2.0 technologies, even more caution is apparent.

The growth of external deployment is important. We're seeing deployment of external blogs and external social networks, though we're lagging with deployment of external crowdsourcing models. This confirms the

a Wikis, blogs, and folksonomies reflect the ability to link data, information, and knowledge previously unlinked (see www.linkeddata.org). Web 2.0 tools "free" users from corporate restrictions on access, content, and transaction processing, so are both a blessing and a curse.

distinction we noted between the internal and external deployment of Web 2.0 technologies during our interviews (see Figure 2).

Table 2 outlines some expectations data. What did senior managers think about the contributions Web 2.0 technologies could make to corporate productivity and management?

The survey data suggests expectations were generally positive, even though most respondents (55%) expect "medium" impact, and 23% expect it to be "high." This combined 78% response suggests the majority of respondents expect the impact of Web 2.0 technologies to be significant. There is a lot of optimism out there.

Table 3 suggests that most respondents expect Web 2.0 technologies to affect knowledge management, collaboration, and communications; many also expected them to positively affect customer relationship management, innovation, and training. Rapid application development was expected to lag relative to the other areas.

Table 4 outlines what happened vs. what respondents thought would happen. For example, knowledge management was expected to be more important than it turned out to be. Collaboration and communications were slightly exaggerated in the expectations survey data, though collaboration and communications were still highly affected by Web 2.0 technologies. Expectations lagged for innovation, training, customer relationship management, and rapid application development. What could explain the optimism that yielded to reality? Cynics might point to pundit hype and vendor exaggeration of technology capabilities, something many vendors do routinely. Others might point to naiveté about early vs. managed-technology adoption processes. Regardless of the reason, we found a gap between what was expected and what actually occurred.

Table 5 shifts to a lower level of analysis, assessing the impact of knowledge management. The four metrics—sharing, retrieving, organizing, and leveraging knowledge—indicate that Web 2.0 technologies contributed significantly to sharing, retrieving, and organizing knowledge but less to leveraging knowledge for problem solving. This makes Web 2.0 technologies (for knowledge management) more descriptive than prescriptive, more operational than strategic.

The impact breakdown is even more interesting. Table 6 suggests that wikis, blogs, and folksonomies/content management lead the way toward improved knowledge management. A surprising finding is the relative lack of impact of RSS filters, because the essence of RSS filtering is knowledge management. Not surprising is that virtual worlds have little impact on knowledge management.

In terms of application development, relatively little ground-up appli-

Table 9. Customer relationship management impact data by ability.

In the area of customer relationship management, have Web 2.0 technologies contributed to your organization's ability to...	Not at all	Very little	Somewhat	A great deal	Response Total
Mine customer data more effectively	42.1% (32)	30.3% (23)	21.1% (16)	6.6% (5)	**76**
"Touch" more customers differently	34.2% (26)	28.9% (22)	22.4% (17)	14.5% (11)	**76**
Solicit customer insights and concerns	36.8% (28)	25.0% (19)	26.3% (20)	11.8% (9)	**76**
Communicate with customers more effectively	32.9% (25)	21.1% (16)	39.5% (30)	6.6% (5)	**76**

Table 10. Web 2.0 technologies and customer relationship management.

In terms of improving customer relationship management, which Web 2.0 technologies have contributed the most? (Please select all that apply.)	Response Percent	Response Total
Wikis	22.4%	17
Internal employee blogs	15.8%	12
External customer blogs	19.7%	15
RSS filters	10.5%	8
Folksonomies/ content management	11.8%	9
Mashups	6.3%	4
Virtual worlds	0%	0
Internal crowdsourcing	1.3%	1
External crowdsourcing	3.9%	3
Internal social networks	9.2%	7
External social networks	17.1%	13
We have not seen any improvement in customer relationship management.	28.9%	22
Other (please specify):	7.9%	6

cation development is going on these days. More and more companies have adapted their processes to those embedded in packaged software applications. Also, a great deal of application development occurs around the customization of functionality extending from packaged applications.

One would think mashup technology would have a dramatic impact on the customization and extension of packaged application-based functionality, an assumption not supported by our survey data. Table 7 suggests a weak relationship across the board between Web 2.0 technologies and application development. This finding also suggests that the new Internet-centered applications architecture may lag as well. While more and more transaction processing occurs outside the corporate firewall, many companies are more comfortable with older application-development enhancement methods and models that do not necessarily involve Web-published application program interfaces, components, and widgets.

Wikis seem to lead the pack of Web 2.0 technologies and their contribution to rapid application development (see Table 8). Wikis apparently represent a suite of new applications companies are deploying. Perhaps surprising is the relatively few survey respondents who view mashups as applications unto themselves or as an applications-development methodology. Web-centric application architectures will use mashup technology extensively to create a new class of applications, though they appear to be more on the drawing board than in the field.

Table 9 indicates that Web 2.0 technologies have had little impact on customer relationship management, a little surprising since several Web 2.0 technologies (such as external customer blogs, wikis, external social networks, and RSS filters) have great potential in this area. This further suggests that we may not be thinking creatively enough about how Web 2.0 technologies can contribute not only to customer relationship management but to other impact areas as well.

Table 10 suggests that wikis and external customer blogs contribute the most to customer relationship management, though, again, the numbers are not compelling. Little confidence was expressed in the use of external social networks. Overall, the data suggests that customer relationship

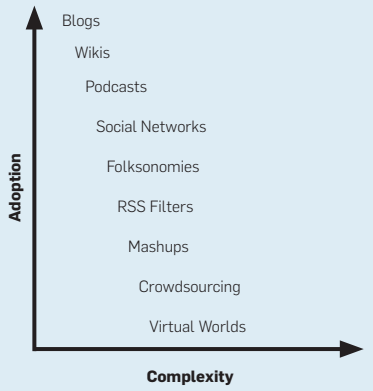
Table 11. Collaboration and communications impact data by ability.

In the area of collaboration and communication, have Web 2.0 technologies contributed to your organization's ability to...	Not at all	Very little	Somewhat	A great deal	Response Total
Coordinate discussions	10.5% (8)	10.5% (8)	55.3% (42)	23.7% (18)	**76**
Reach more people faster	3.9% (3)	17.1% (13)	50.0% (38)	28.9% (22)	**76**
Synchronize projects and tasks	13.2% (10)	22.4% (17)	56.6% (43)	7.9% (6)	**76**
Audit communications streams	30.3% (23)	31.6% (24)	32.9% (25)	5.3% (4)	**76**

Table 12. Web 2.0 technologies, collaboration, communication.

In terms of improving rapid application development, which Web 2.0 technologies have contributed the most? (Please select all that apply.)	Response Percent	Response Total
Wikis	67.1%	51
Internal employee blogs	42.1%	32
External customer blogs	11.8%	9
RSS filters	17.1%	13
Folksonomies/ content management	18.4%	14
Mashups	5.3%	4
Virtual worlds	2.6%	2
Internal crowdsourcing	6.6%	5
External crowdsourcing	1.3%	1
Internal social networks	25.0%	19
External social networks	13.2%	10
We have not seen any improvement in rapid application development.	9.2%	7
Other (please specify):	3.9%	3

Figure 3. Adoption and complexity.



management is not viewed as a prime impact area for Web 2.0 technologies, though this attitude might change over time.

Table 11 shifts the focus to collaboration and communication, where, as expected, the impact is significant. Wikis are the runaway hit, followed by blogs and external social networks. However, we found a lower level of deployment sophistication than the ideal. For example, the "auditing" of communications and collaboration streams (classic business intelligence) lags well behind other impact areas. The power of many Web 2.0 technologies often involves the ability to perform primary and secondary analyses of transactions, communications patterns, and customer service. Our survey data appears to indicate that we're seeing a toe-in-the-water effect, where companies experiment with initial deployments but stop short of full commitment through total exploitation of the technologies.

Table 12 confirms all this, with wikis, internal blogs, and internal social networks leading the way in collaboration and communications. While this trend is to be expected, many other opportunities have yet to be exploited. Table 12 also suggests weakness in externally focused Web 2.0 technology deployment—surprising in light of the technology's capabilities. We can infer from this data that external applications lag internal ones and that over time significant collaboration and communication applications can be expected. Why such optimism? Because Web 2.0 technology capabilities are essentially built on ubiquitous collaboration and communication.

Table 13 turns to innovation, though there's not much enthusiasm here, despite enough progress to excite those who think Web 2.0 technology can eventually contribute to innovation. Crowdsourcing is an especially powerful Web 2.0 innovation technology, along with RSS filters, wikis, and blogs.

Table 14 outlines how Web 2.0 technologies contribute to innovation. Very surprising is the relative unimportance survey respondents ascribe to external crowdsourcing. (Does anyone believe virtual worlds are useful for anything?)

Training is the final area we assessed. Table 15 suggests that survey respondents have not yet defined how Web 2.0 technologies can contribute to training. While wikis are natural-

Figure 4. Segmentation of Web 2.0 technologies.

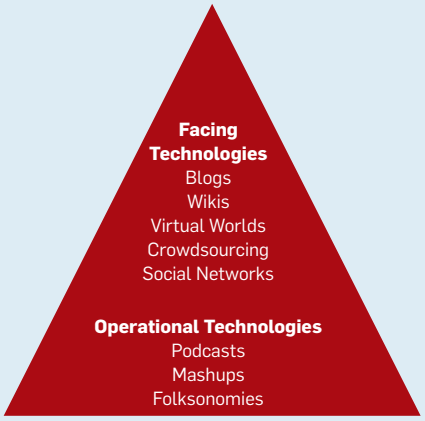


Table 13. Innovation impact data by ability.

In the area of innovation, have Web 2.0 technologies contributed to your organization's ability to...	Not at all	Very little	Somewhat	A great deal	Response Total
Organize innovation	27.6% (21)	22.4% (17)	39.5% (30)	10.5% (8)	**76**
Improve R&D success	36.8% (28)	15.8% (12)	35.5% (27)	11.8% (9)	**76**
Increase the number of innovation initiatives	35.5% (27)	19.7% (15)	31.6% (24)	13.2% (10)	**76**
Productize innovations more effectively	39.5% (30)	14.5% (11)	38.2% (29)	7.9% (6)	**76**

Table 14. Web 2.0 technologies and innovation.

In terms of improving innovation, which Web 2.0 technologies have contributed the most? (Please select all that apply.)	Response Percent	Response Total
Wikis	50.0%	38
Internal employee blogs	30.3%	23
External customer blogs	9.2%	7
RSS filters	9.2%	7
Folksonomies/ content management	10.5%	8
Mashups	5.3%	4
Virtual worlds	1.3%	1
Internal crowdsourcing	7.9%	6
External crowdsourcing	3.9%	3
Internal social networks	17.1%	13
External social networks	5.3%	4
We have not seen any improvement in customer relationship management.	26.3%	20
Other (please specify):	3.9%	3

born trainers, Web 2.0 technologies can contribute much more. What were the respondents missing? Table 16 provides the details. While wikis "win," other technologies are discounted, at least for now. Meanwhile, this is where virtual worlds might actually contribute to education and learning, though there's not much evidence to suggest that anyone agrees.

Interpretation

What did we learn from the interviews, observations, and survey? Security remains a major issue in the adoption of Web 2.0 technology. Beyond it, there's also internal control and prudence versus flexibility, even liability. Some companies block access to social networking sites from corporate networks; others are creating their own corporate social networking sites, though we found companies concerned about the amount of time employees spend on them.

Our interview, observation, and survey data all suggest the lowest-hanging fruit is—surprise!—picked first. Wikis, blogs, and social networks, perhaps due to their consumer-to-consumer origins, have been deployed more than the other Web 2.0 technologies. Fear of the unknown might explain why virtual worlds, folksonomies, crowdsourcing, and even RSS filters have lagged deployment of the wiki/blog/social network big three.

It also appears the survey respondents have not yet discovered the second-level potential of Web 2.0 technologies. Mashup technology is potentially extremely powerful but has not yet penetrated the rapid-application-development mind-set. Similarly, the customer-relationship-management mind-set is under-influenced by Web 2.0 technologies.

One important factor constraining adoption of Web 2.0 technology is the existing applications portfolio in companies with substantial technology budgets. In addition to the perennial issues around asset amortization, not-invented-here constraints restrict introduction of new applications based on new technologies. This walled-garden effect is real in many companies, restricting adoption of new technologies, applications, and even processes.

Some Web 2.0 technologies are operational, and some are employee- and customer-facing. Figures 3 and 4 suggest a relationship between complexity and adoption and an important distinction between operational and facing technologies. We should assume that simple (versus complex) facing technologies will be adopted more quickly than complicated operational ones.

Web 2.0 technology also fuels the broad area of information warfare. Just as cyberbullying is a nasty trend in the consumer world, anonymous blogging can hurt business, images, and brands. The number of incidents designed to harm companies (sometimes specifically targeted) is growing dramatically. Companies will have to increase their cybervigilance and invest in countermeasures. Web 2.0 technology also empowers disgruntled employees who might want to hurt their companies. Whistleblow-

Table 15. Training impact data by ability.

In the area of training, have Web 2.0 technologies contributed to your organization's ability to...	Not at all	Very little	Somewhat	A great deal	Response Total
Support traditional training	44.7% (34)	22.4% (17)	26.3% (20)	6.6% (5)	**76**
Modify and evolve training content	36.8% (28)	18.7% (15)	30.3% (23)	12.2% (10)	**76**
Suppport distance training	34.2% (26)	21.1% (16)	27.6% (21)	17.1% (13)	**76**
Distribute training content	35.5% (27)	19.7% (15)	35.5% (27)	9.2% (7)	**76**

Table 16. Web 2.0 technologies and training.

In terms of improving training, which Web 2.0 technologies have contributed the most? (Please select all that apply.)	Response Percent	Response Total
Wikis	40.8%	31
Internal employee blogs	21.1%	16
External customer blogs	10.5%	8
RSS filters	11.8%	9
Folksonomies/ content management	14.5%	11
Mashups	3.9%	3
Virtual worlds	2.6%	2
Internal crowdsourcing	2.6%	2
External crowdsourcing	0%	0
Internal social networks	14.5%	11
External social networks	5.3%	4
We have not seen any improvement in rapid application development.	28.9%	22
Other (please specify):	9.2%	7

ing promises to take on new forms through Web 2.0 channels.

As more Web 2.0 technologies are deployed, and as early impact is positively assessed, additional deployment and additional productivity can be expected. Momentum breeds momentum, and the second-order impact of the technologies will be felt as momentum grows. While "simple is good" today, "complex and powerful" will define tomorrow's deployment of Web 2.0 and 3.0 technologies.

Web 3.0 technologies should be anticipated. According to Wikipedia.org, Web 3.0 technologies include: "The emergence of 'The Data Web' as structured data records are published to the Web in reusable and remotely queryable formats. The Data Web enables a new level of data integration and application interoperability, making data as openly accessible and linkable as Web pages. The Data Web is the first step on the path toward the full Semantic Web. The full Semantic Web will widen the scope such that both structured data and even what is traditionally thought of as unstructured or semi-structured content (such as Web pages and documents) will be widely available in RDF and OWL semantic formats. Web site parse templates will be used by Web 3.0 crawlers to get more precise information about Web sites' structured content. Web 3.0 has also been used to describe an evolutionary path for the Web that leads to artificial intelligence that can reason about the Web in a quasi-human fashion."

Next-generation Web technology will be proactive, intelligent, contextual, automated, and adaptive. While we examined adoption of Web 2.0 technologies, imagine the analyses of Web 3.0 technology adoption we'll eventually conduct. When technology integrates seamlessly into business processes at all levels we can expect impact to be immediate and dramatic. The full potential of Web 3.0 is years away, but the drivers of Web 2.0 technology adoption already provide clues to how ubiquitous Web 3.0 is likely to be.

Regardless of the reason, we found a gap between what was expected and what actually occurred.

Acknowledgments

I would like to thank the Alfred P. Sloan Foundation for supporting the interview and direct observation processes; Villanova University and the Cutter Consortium for supporting the collection of the survey data; and A. Frank Mayadas of the Alfred P. Sloan Foundation for his always excellent comments and insights along the way.

References

1. Ahn, Y.-Y., Han, S., Kwak, H., Moon, S., and Jeong, H. Semantic Web and Web 2.0: Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th International Conference on World Wide Web* (Calgary, 2007).
2. Boll, S. MultiTube: Where Web 2.0 and multimedia could meet. *IEEE MultiMedia 14,* 1 (Jan. 2007).
3. Brier, J. Guidelines: Web accessibility highlights and trends. In *Proceedings of the 2004 International Cross-Disciplinary Workshop on Web Accessibility* (Manchester, England, 2004).
4. Fox, G. Implications of Web 2.0 for the semantic grid. In *Proceedings of the Second International Conference on Semantics, Knowledge, and Grid* (Guilin, Guangxi, China 2006).
5. Jaokar, A. and Fish, T. *Mobile Web 2.0: The Innovator's Guide to Developing and Marketing Next-Generation Wireless/Mobile Applications.* Futuretext, London, Aug. 2006.
6. Lin, K.-J. Building Web 2.0. *IEEE Computer 40,* 5 (May 2007).
7. Lin, K.-J. Serving Web 2.0 with SOA: Providing the technology for innovation and specialization. In *Proceedings of the IEEE International Conference on e-Business Engineering* (Los Angeles, 2006).
8. Losinski, R. Patrolling Web 2.0. *THE (Technological Horizons in Education) Journal 34.*
9. Mahmood, O. Developing Web 2.0 applications for semantic Web of trust. In *Proceedings of the International Conference on Information Technology,* 2007.
10. Majchrzak, A., Wagner, C., and Yeates, D. Corporate wiki users: Results of a survey. In *Proceedings of WikiSym,* 2006.
11. McKinsey & Co. *How Businesses Are Using Web 2.0: A McKinsey Global Survey,* 2007.
12. Minol, K., Spelsberg, G., Schulte, E., and Morris, N. Portals, blogs and co.: The role of the Internet as a medium of science communication. *Biotechnology Journal 2,* 8 (Aug. 2007).
13. Mori, M., Miura, T., and Shioya, I. Topic detection and tracking for news Web pages. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence,* 2006.
14. Orr, B. Parsing the Meaning of Web 2.0. *ABA Banking Journal 9.*
15. Poer, C. and Petrie, H. Accessibility and guidelines: Accessibility in non-professional Web authoring tools: A missed Web 2.0 opportunity? In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility,* 2007.
16. Tredinnick, L. Web 2.0 and business. *Business Information Review 23,* 4 (2006), 228–234.
17. van der Vlist, E., Ayers, D., Bruchez, E., Fawcett, J., and Vernet, A. *Professional Web 2.0 Programming.* Wrox Professional Guides, Wrox Press Ltd., Nov. 2006.
18. Wagner, C. and Majchrzak, A. Enabling customer centricity using wikis and the wiki way. *Journal of Management Information Systems 23,* 3 (2007).
19. Yanbe, Y., Jatowt, A., Nakamura, S., and Tanaka, K. Social networks: Can social bookmarking enhance search on the Web? In *Proceedings of the Conference on Digital Libraries,* 2007.
20. Zajicek, M. Web 2.0: Hype or happiness? In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility,* 2007.

Stephen J. Andriole (stephen.andriole@villanova.edu) is the Thomas G. Labrecque Professor of Business in the Department of Management & Operations in the Villanova School of Business at Villanova University, Villanova, PA.

DOI:10.1145/1859204.1859227

What are Bayesian networks and why are their applications growing across all fields?

BY ADNAN DARWICHE

Bayesian Networks

BAYESIAN NETWORKS HAVE been receiving considerable attention over the last few decades from scientists and engineers across a number of fields, including computer science, cognitive science, statistics, and philosophy. In computer science, the development of Bayesian networks was driven by research in artificial intelligence, which aimed at producing a practical framework for commonsense reasoning.[29] Statisticians have also contributed to the development of Bayesian networks, where they are studied under the broader umbrella of probabilistic graphical models.[5,11]

Interestingly enough, a number of other more specialized fields, such as genetic linkage analysis, speech recognition, information theory and reliability analysis, have developed representations that can be thought of as concrete instantiations or restricted cases of Bayesian networks. For example, pedigrees and their associated phenotype/genotype information, reliability block diagrams, and hidden Markov models (used in many fields including speech recognition and bioinformatics) can all be viewed as Bayesian networks. Canonical instances of Bayesian networks also exist and have been used to solve standard problems that span across domains such as computer vision, the Web, and medical diagnosis.

So what are Bayesian networks, and why are they widely used, either directly or indirectly, across so many fields and application areas? Intuitively, Bayesian networks provide a systematic and localized method for structuring probabilistic information about a situation into a coherent whole. They also provide a suite of algorithms that allow one to automatically derive many implications of this information, which can form the basis for important conclusions and decisions about the corresponding situation (for example, computing the overall reliability of a system, finding the most likely message that was sent across a noisy channel, identifying the most likely users that would respond to an ad, restoring a noisy image, mapping genes onto a chromosome, among others). Technically speaking, a Bayesian network is a compact representation of a probability distribution that is usually too large to be handled using traditional specifications from probability and statistics such as tables and equations. For example, Bayesian networks with thousands of variables have been constructed and reasoned about successfully, allowing one to efficiently represent and reason about probability distributions whose size is exponential in that number of variables (for example, in genetic link-

» key insights

- **Bayesian networks provide a systematic and localized method for structuring probabilistic information about a situation into a coherent whole, and are supported by a suite of inference algorithms.**
- **Bayesian networks have been established as a ubiquitous tool for modeling and reasoning under uncertainty.**
- **Many applications can be reduced to Bayesian network inference, allowing one to capitalize on Bayesian network algorithms instead of having to invent specialized algorithms for each new application.**

age analysis,[12] low-level vision,[34] and networks synthesized from relational models[4]).

For a concrete feel of Bayesian networks, Figure 1 depicts a small network over six binary variables. Every Bayesian network has two components: a directed acyclic graph (called a structure), and a set of conditional probability tables (CPTs). The nodes of a structure correspond to the variables of interest, and its edges have a formal interpretation in terms of probabilistic independence. We will discuss this interpretation later, but suffice to say here that in many practical applications, one can often interpret network edges as signifying direct causal influences. A Bayesian network must include a CPT for each variable, which quantifies the relationship between that variable and its parents in the network. For example, the CPT for variable *A* specifies the conditional probability distribution of *A* given its parents *F* and *T*. According to this CPT, the probability of *A* = `true` given *F* = `true` and *T* = `false` is *Pr*(*A*=`true`|*F* = `true`; *T* = `false`) = .9900 and is called a network *parameter*.[a]

A main feature of Bayesian networks is their guaranteed consistency and completeness as there is one and only one probability distribution that satisfies the constraints of a Bayesian network. For example, the network in Figure 1 induces a unique probability distribution over the 64 instantiations of its variables. This distribution provides enough information to attribute a probability to every event that can be expressed using the variables appearing in this network, for example, the probability of alarm tampering given no smoke and a report of people leaving the building.

Another feature of Bayesian networks is the existence of efficient algorithms for computing such probabilities without having to explic-

a Bayesian networks may contain continuous variables, yet our discussion here is restricted to the discrete case.

IMAGE BY ROBERT HODGIN

itly generate the underlying probability distribution (which would be computationally infeasible for many interesting networks). These algorithms, to be discussed in detail later, apply to any Bayesian network, regardless of its topology. Yet, the efficiency of these algorithms—and their accuracy in the case of approximation algorithms—may be quite sensitive to this topology and the specific query at hand. Interestingly enough, in domains such as genetics, reliability analysis, and information theory, scientists have developed algorithms that are indeed subsumed by the more general algorithms for Bayesian networks. In fact, one of the main objectives of this article is to raise awareness about these connections. The more general objective, however, is to provide an accessible introduction to Bayesian networks, which allows scientists and engineers to more easily identify problems that can be reduced to Bayesian network inference, putting them in a position where they can capitalize on the vast progress that has been made in this area over the last few decades.

Causality and Independence

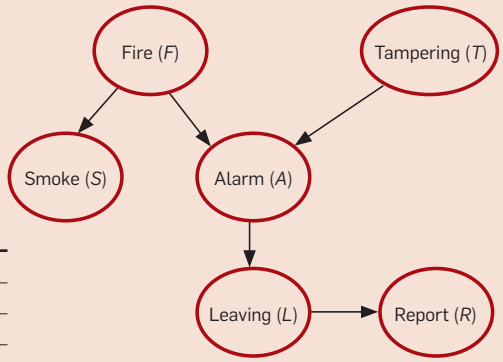
We will start by unveiling the central insight behind Bayesian networks that allows them to compactly represent very large distributions. Consider Figure 1 and the associated CPTs. Each probability that appears in one of these CPTs does specify a constraint that must be satisfied by the distribution induced by the network. For example, the distribution must assign the probability .01 to having smoke without fire, $Pr(S = \texttt{true} | F = \texttt{false})$, since this is specified by the CPT of variable S. These constraints, however, are not sufficient to pin down a unique probability distribution. So what additional information is being appealed to here?

The answer lies in the structure of a Bayesian network, which specifies additional constraints in the form of probabilistic conditional independencies. In particular, every variable in the structure is assumed to become independent of its non-descendants once its parents are known. In Figure 1, variable L is assumed to become independent of its non-descendants T, F, S once its parent A is known. In other words, once the value of variable A is known, the probability distribution of variable L will no longer change due to new information about variables T, F and S. Another example from Figure 1: variable A is assumed to become independent of its non-descendant S once its parents F and T are known. These independence constraints are known as the Markovian assumptions of a Bayesian network. Together with the numerical constraints specified by CPTs, they are satisfied by exactly one probability distribution.

Does this mean that every time a Bayesian network is constructed, one must verify the conditional independencies asserted by its structure? This really depends on the construction method. I will discuss three main methods in the section entitled "How Are Bayesian Networks Constructed?" that include subjective construction, synthesis from other specifications, and learning from data. The first method is the least systematic, but even in that case, one rarely thinks about conditional independence when constructing networks. Instead, one thinks about causality, adding the edge $X \rightarrow Y$ whenever X is perceived to be a direct cause of Y. This leads to a causal structure in which the Markovian assumptions read: each variable becomes independent of its non-effects once its direct causes are known. The ubiquity of Bayesian networks stems from the fact that people are quite good at identifying direct causes from a given set of variables, and at deciding whether the set of variables contains all of the relevant direct causes. This ability is all that one needs for constructing a causal structure.

The distribution induced by a Bayesian network typically satisfies additional independencies, beyond the Markovian ones discussed above. Moreover, all such independencies can be identified efficiently using a graphical test known as *d-separation*.[29] According to this test, variables X and

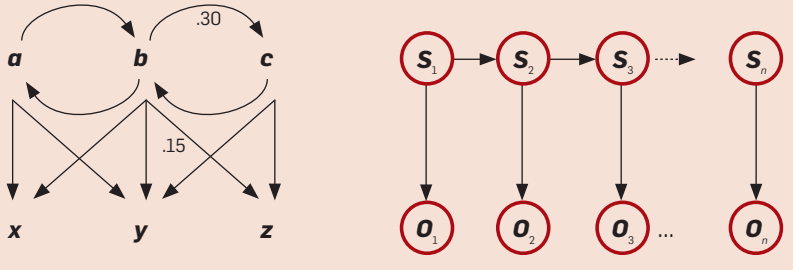
Figure 1. A Bayesian network with some of its conditional probability tables (CPTs).



Fire	Smoke	$\theta_{s|f}$
true	true	.90
false	true	.01

Fire	Tampering	Alarm	$\theta_{a|f,t}$
true	true	true	.5000
true	false	true	.9900
false	true	true	.8500
false	false	true	.0001

Figure 2. A hidden Markov model (HMM) and its corresponding dynamic Bayesian network (DBN).



(a) HMM (state diagram)

(b) DBN

Y are guaranteed to be independent given variables Z if every path between X and Y is blocked by Z. Intuitively, a path is blocked when it cannot be used to justify a dependence between X and Y in light of our knowledge of Z. For an example, consider the path $\alpha : S \leftarrow F \rightarrow A \leftarrow T$ in Figure 1 and suppose we know the alarm has triggered (that is, we know the value of variable A). This path can then be used to establish a dependence between variables S and T as follows. First, observing smoke increases the likelihood of fire since fire is a direct cause of smoke according to path α. Moreover, the increased likelihood of fire explains away tampering as a cause of the alarm, leading to a decrease in the probability of tampering (fire and tampering are two competing causes of the alarm according to path α). Hence, the path could be used to establish a dependence between S and T in this case. Variables S and T are therefore not independent given A due to the presence of this unblocked path. One can verify, however, that this path cannot be used to establish a dependence between S and T in case we know the value of variable F instead of A. Hence, the path is blocked by F.

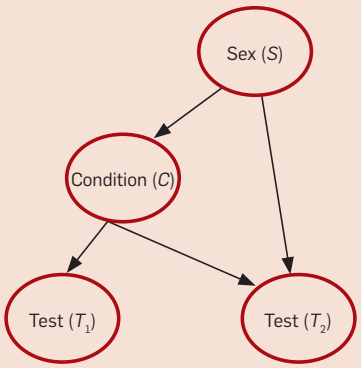
Even though we appealed to the notion of causality when describing the d-separation test, one can phrase and prove the test without any appeal to causality—we only need the Markovian assumptions. The full d-separation test gives the precise conditions under which a path between two variables is blocked, guaranteeing independence whenever all paths are blocked. The test can be implemented in time linear in the Bayesian network structure, without the need to explicitly enumerate paths as suggested previously.

The d-separation test can be used to directly derive results that have been proven for specialized probabilistic models used in a variety of fields. One example is hidden Markov models (HMMs), which are used to model dynamic systems whose states are not observable, yet their outputs are. One uses an HMM when interested in making inferences about these changing states, given the sequence of outputs they generate. HMMs are widely used in applications requiring temporal pattern recognition, including speech, handwriting, and gesture recognition; and various problems in bioinformatics.[31] Figure 2a depicts an HMM, which models a system with three states (a, b, c) and three outputs (x, y, z). The figure depicts the possible transitions between the system states, which need to be annotated by their probabilities. For example, state b can transition to states a or c, with a 30% chance of transitioning to state c. Each state can emit a number of observable outputs, again, with some probabilities. In this example, state b can emit any of the three outputs, with output z having a 15% chance of being emitted by this state.

This HMM can be represented by the Bayesian network in Figure 2b.[32] Here, variable S_t has three values a, b, c and represents the system state at time t, while variable O_t has the values x, y, z and represents the system output at time t. Using d-separation on this network, one can immediately derive the characteristic property of HMMs: once the state of the system at time t is known, its states and outputs at times $> t$ become independent of its states and outputs at times $< t$.

We also note the network in Figure 2b is one of the simplest instances of what is known as *dynamic Bayesian networks* (DBNs).[9] A number of extensions have been considered for HMMs, which can be viewed as more structured instances of DBNs. When proposing such extensions, however, one has the obligation of offering a corresponding algorithmic toolbox for inference. By viewing these extended HMMs as instances of Bayesian networks, however, one immediately inherits the corresponding Bayesian network algorithms for this purpose.

Figure 3. A Bayesian network that models a population, a medical condition, and two corresponding tests.



C = yes means that the condition C is present, and T_i = +ve means the result of test T_i is positive.

S	θ_s
male	.55

S	C	$\theta_{c|s}$
male	yes	.05
female	yes	.01

C	T_1	$\theta_{t_1|c}$
yes	+ve	.80
no	+ve	.20

S	C	T_2	$\theta_{t_2|s,c}$
male	yes	+ve	.80
male	no	+ve	.20
female	yes	+ve	.95
female	no	+ve	.05

How are Bayesian Networks Constructed?

One can identify three main methods for constructing Bayesian networks.[8] According to the first method, which is largely subjective, one reflects on their own knowledge or the knowledge of others (typically, perceptions about causal influences) and then captures them into a Bayesian network. The network in Figure 1 is an example of this construction method. The network structure of Figure 3 depicts another example, yet the parameters of this network can be obtained from more formal sources, such as population statistics and test specifications. According to this network, we have a population that is 55% males and 45% females, whose members can suffer from a medical condition C that is more likely to occur in males. Moreover, two diagnostic tests are available for detecting this condition, T_1 and T_2, with the second test being more effective on females. The CPTs of this network also reveal that the two tests are equally effective on males.

The second method for constructing Bayesian networks is based on automatically synthesizing them from some other type of formal knowledge.

Figure 4. A reliability block diagram (top), with a systematic method for mapping its blocks into Bayesian network fragments (bottom).

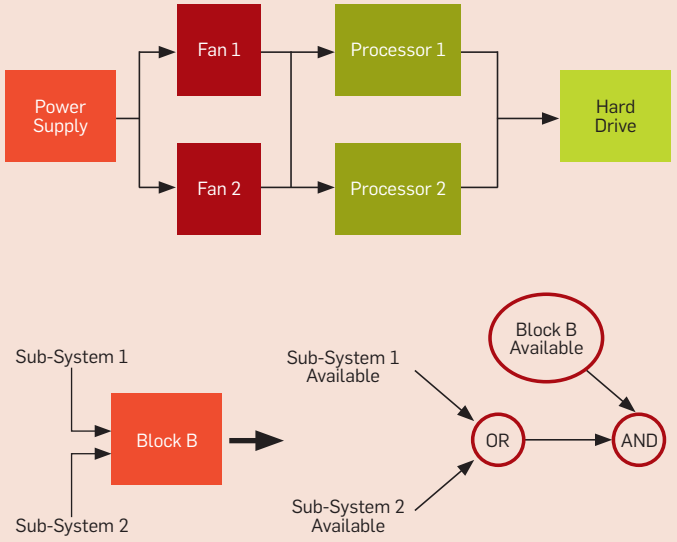
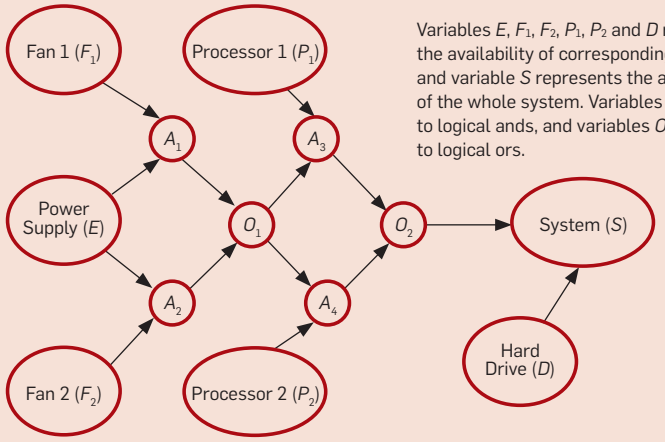


Figure 5. A Bayesian network generated automatically from a reliability block diagram.



For example, in many applications that involve system analysis, such as reliability and diagnosis, one can synthesize a Bayesian network automatically from a formal system design. Figure 4 depicts a reliability block diagram (RBD) used in reliability analysis. The RBD depicts system components and the dependencies between their availability. For example, Processor 1 requires either of the fans for its availability, and each of the fans requires power for its availability. The goal here is to compute the overall reliability of the system (probability of its availability) given the reliabilities of each of its components. Figure 4 shows also how one may systematically convert each block in an RBD into a Bayesian network fragment, while Figure 5 depicts the corresponding Bayesian network constructed according to this method. The CPTs of this figure can be completely constructed based on the reliabilities of individual components (not shown here) and the semantics of the transformation method.[8]

The third method for constructing Bayesian networks is based on learning them from data, such as medical records or student admissions data. Consider Figure 3 and the data set depicted in the table here as an example.

Sex S	Condition C	Test T_1	Test T_2
male	no	?	–ve
male	?	–ve	+ve
female	yes	+ve	?
⋮	⋮	⋮	⋮

Each row of the table corresponds to an individual and what we know about them. One can use such a data set to learn the network parameters given its structure, or learn both the structure and its parameters. Learning parameters only is an easier task computationally. Moreover, learning either structure or parameters always becomes easier when the data set is complete—that is, the value of each variable is known in each data record.

Since learning is an inductive process, one needs a principle of induction to guide the learning process. The two main principles for this purpose lead to the *maximum likelihood* and *Bayesian* approaches to learning (see, for example, the work of [5,8,17,22,27]). The maximum likelihood approach favors Bayesian networks that maximize the probability of observing the given data set. The Bayesian approach on the other hand uses the likelihood principle in addition to some prior information which encodes preferences on Bayesian networks.

Suppose we are only learning network parameters. The Bayesian approach allows one to put a prior distribution on the possible values of each network parameter. This prior distribution, together with the data set, induces a posterior distribution on the values of that parameter. One can then use this posterior to pick a value for that parameter (for example, the distribution mean). Alternatively, one can decide to avoid committing to a fixed parameter value, while computing answers to queries by averaging over all possible parameter values according to their posterior probabilities.

It is critical to observe here that the term "Bayesian network" does not necessarily imply a commitment to the Bayesian approach for learning networks. This term was coined by Judea Pearl[28] to emphasize three aspects: the often subjective nature of the information used in constructing these networks; the reliance on Bayes condi-

tioning when reasoning with Bayesian networks; and the ability to perform causal as well as evidential reasoning on these networks, which is a distinction underscored by Thomas Bayes.[1]

These learning approaches are meant to induce Bayesian networks that are meaningful independently of the tasks for which they are intended. Consider for example a network which models a set of diseases and a corresponding set of symptoms. This network may be used to perform *diagnostic* tasks, by inferring the most likely disease given a set of observed symptoms. It may also be used for *prediction* tasks, where we infer the most likely symptom given some diseases. If we concern ourselves with only one of these tasks, say diagnostics, we can use a more specialized induction principle that optimizes the diagnostic performance of the learned network. In machine learning jargon, we say we are learning a *discriminative model* in this case, as it is often used to discriminate among patients according to a predefined set of classes (for example, has cancer or not). This is to be contrasted with learning a *generative model*, which is to be evaluated based on its ability to generate the given data set, regardless of how it performs on any particular task.

We finally note that it is not uncommon to assume some canonical network structure when learning Bayesian networks from data, in order to reduce the problem of learning structure and parameters to the easier problem of learning parameters only. Perhaps the most common such structure is what is known as naïve Bayes: $C \rightarrow A_1, \ldots, C \rightarrow A_n$, where C is known as the class variable and variables $A_1,\ldots,A_n$ are known as attributes. This structure has proven to be very popular and effective in a number of applications, in particular classification and clustering.[14]

Canonical Bayesian Networks

A number of canonical Bayesian networks have been proposed for modeling some well-known problems in a variety of fields. For example, genetic linkage analysis is concerned with mapping genes onto a chromosome, utilizing the fact that the distance between genes is inversely proportional to the extent to which genes are linked (two genes are linked when it is more likely than not that their states are passed together from a single grandparent, instead of one state from each grandparent). To assess the likelihood of a linkage hypothesis, one uses a pedigree with some information about the genotype and phenotype of associated individuals. Such information can be systematically translated into a Bayesian network (see Figure 6), where the likelihood of a linkage hypothesis corresponds to computing the probability of an event with respect to this network.[12] By casting this problem in terms of inference on Bayesian networks, and by capitalizing on the state-of-the-art algorithms for this purpose, the scalability of genetic linkage analysis was advanced considerably, leading to the most efficient algorithms for exact linkage computations on general pedigrees (for example, the SUPER-

Figure 6. A Bayesian network generated automatically from a pedigree that contains three individuals.

Variables GP_{ij} and GM_{ij} represent the states of gene j for individual i, which are inherited from the father and mother of i, respectively. Variable P_{ij} represents the phenotype determined by gene j of individual i. Variables SP_{ij} and SM_{ij} are meant to model recombination events, which occur when the states of two genes inherited from a parent do not come from a single grandparent. The CPTs of these variables contain parameters which encode linkage hypotheses.

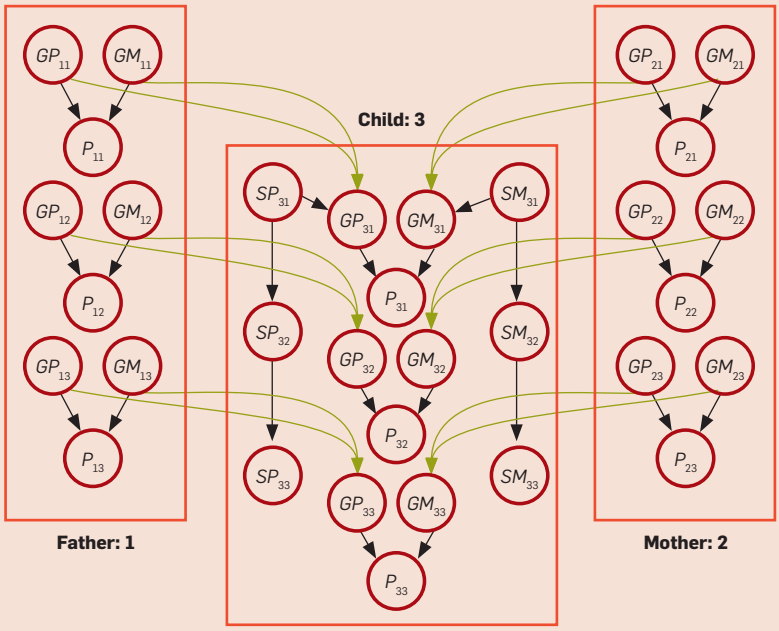
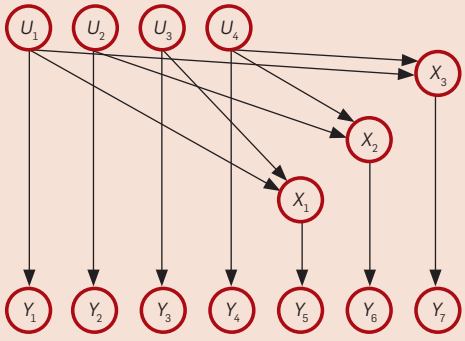


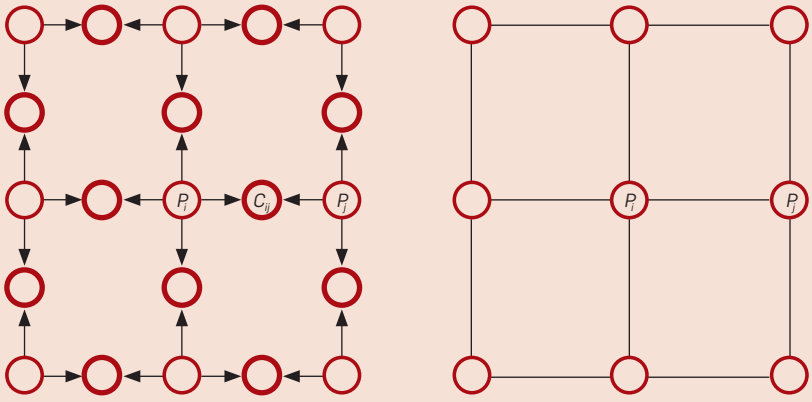
Figure 7. A Bayesian network that models a noisy channel with input ($U_1,\ldots,U_4, X_1,\ldots,X_3$) and output ($Y_1,\ldots,Y_7$).



Bits $X_1,\ldots,X_3$ are redundant coding bits and their CPTs capture the coding protocol. The CPTs of variables Y_i capture the noise model. This network can be used to recover the most likely sent message $U_1,\ldots,U_4$ given a received message $Y_1,\ldots,Y_7$.

Figure 8. Images from left to right: input, restored (using Bayesian network inference) and original.

Figure 9. Modeling low-level vision problems using two types of graphical models: Bayesian networks and MRFs.



(a) Bayesian network

(b) Markov random field (MRF)

LINK program initiated by Fishelson and Geiger[12]).

Canonical models also exist for modeling the problem of passing information over a noisy channel, where the goal here is to compute the most likely message sent over such a channel, given the channel output.[13] For example, Figure 7 depicts a Bayesian network corresponding to a situation where seven bits are sent across a noisy channel (four original bits and three redundant ones).

Another class of canonical Bayesian networks has been used in various problems relating to vision, including image restoration and stereo vision. Figure 8 depicts two examples of images that were restored by posing a query to a corresponding Bayesian network. Figure 9a depicts the Bayesian network in this case, where we have one node P_i for each pixel i in the image—the values p_i of P_i represent the gray levels of pixel i. For each pair of neighboring pixels, i and j, a child node C_{ij} is added with a CPT that imposes a smoothness constraint between the pixels. That is, the probability $Pr(C_{ij} = \texttt{true} \mid P_i = p_i, P_j = p_j)$ specified by the CPT decreases as the difference in gray levels $| p_i - p_j |$ increases. The only additional information needed to completely specify the Bayesian network is a CPT for each node P_i, which provides a prior distribution on the gray levels of each pixel i. These CPTs are chosen to give the highest probability to the gray level v_i appearing in the input image, with the prior probability $Pr(P_i = p_i)$ decreasing as the difference $| p_i - v_i |$ increases. By simply adjusting the domain and the prior probabilities of nodes P_i, while asserting an appropriate degree of smoothness using variables C_{ij}, one can use this model to perform other "pixel-labeling" tasks such as stereo vision, photomontage, and binary segmentation.[34] The formulation of these tasks as inference on Bayesian networks is not only elegant, but has also proven to be very powerful. For example, such inference is the basis for almost all top-performing stereo methods.[34]

Canonical Bayesian network models have also been emerging in recent years in the context of other important applications, such as the analysis of documents, and text. Many of these networks are based on topic models that view a document as arising from a set of unknown topics, and provide a framework for reasoning about the connections between words, documents, and underlying topics.[2,33] Topic models have been applied to many kinds of documents, including email, scientific abstracts, and newspaper archives, allowing one to utilize inference on Bayesian networks to tackle tasks such as measuring document similarity, discovering emergent topics, and browsing through documents based on their underlying content instead of traditional indexing schemes.

What Can One Do with a Bayesian Network?

Similar to any modeling language, the value of Bayesian networks is mainly tied to the class of queries they support.

Consider the network in Figure 3 for an example and the following queries: Given a male that came out positive on both tests, what is the probability he has the condition? Which group of the population is most likely to test negative on both tests? Considering the network in Figure 5: What is the overall reliability of the given system? What is the most likely configuration

of the two fans given that the system is unavailable? What single component can be replaced to increase the overall system reliability by 5%? Consider Figure 7: What is the most likely channel input that would yield the channel output 1001100? These are example questions that would be of interest in these application domains, and they are questions that can be answered systematically using three canonical Bayesian network queries.[8] A main benefit of using Bayesian networks in these application areas is the ability to capitalize on existing algorithms for these queries, instead of having to invent a corresponding specialized algorithm for each application area.

Probability of Evidence. This query computes the probability $Pr(\mathbf{e})$, where $\mathbf{e}$ is an assignment of values to some variables $\mathbf{E}$ in the Bayesian network—$\mathbf{e}$ is called a variable instantiation or evidence in this case. For example, in Figure 3, we can compute the probability that an individual will come out positive on both tests using the probability-of-evidence query $Pr(T_1 = +\texttt{ve}, T_2 = +\texttt{ve})$. We can also use the same query to compute the overall reliability of the system in Figure 5, $Pr(S = \texttt{avail})$. The decision version of this query is known to be *PP*-complete. It is also related to another common query, which asks for computing the probability $Pr(x|\mathbf{e})$ for each network variable X and each of its values x. This is known as the *node marginals* query.

Most Probable Explanation (MPE). Given an instantiation $\mathbf{e}$ of some variables $\mathbf{E}$ in the Bayesian network, this query identifies the instantiation $\mathbf{q}$ of variables $\mathbf{Q} = \overline{\mathbf{E}}$ that maximizes the probability $Pr(\mathbf{q}|\mathbf{e})$. In Figure 3, we can use an MPE query to find the most likely group, dissected by sex and condition, that will yield negative results for both tests, by taking the evidence $\mathbf{e}$ to be $T_1 = -\text{ve}; T_2 = -\text{ve}$ and $\mathbf{Q} = \{S, C\}$. We can also use an MPE query to restore images as shown in Figures 8 and 9. Here, we take the evidence $\mathbf{e}$ to be $C_{ij} = \texttt{true}$ for all i, j and $\mathbf{Q}$ to include P_i for all i. The decision version of MPE is known to be *NP*-complete and is therefore easier than the probability-of-evidence query under standard assumptions of complexity theory.

Maximum a Posteriori Hypothesis (MAP). Given an instantiation $\mathbf{e}$ of some

Figure 10. An arithmetic circuit for the Bayesian network $B \leftarrow A \rightarrow C$. Inputs labeled with θ variables correspond to network parameters, while those labeled with λ variables capture evidence.

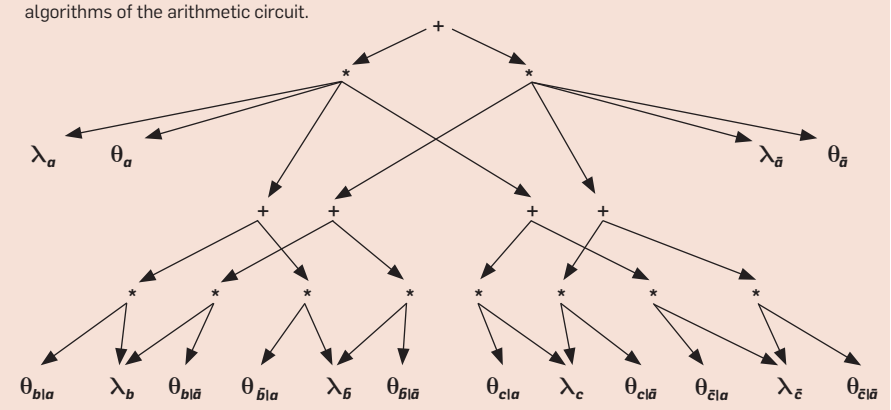


Figure 11. Two networks that represent the same set of conditional independencies.

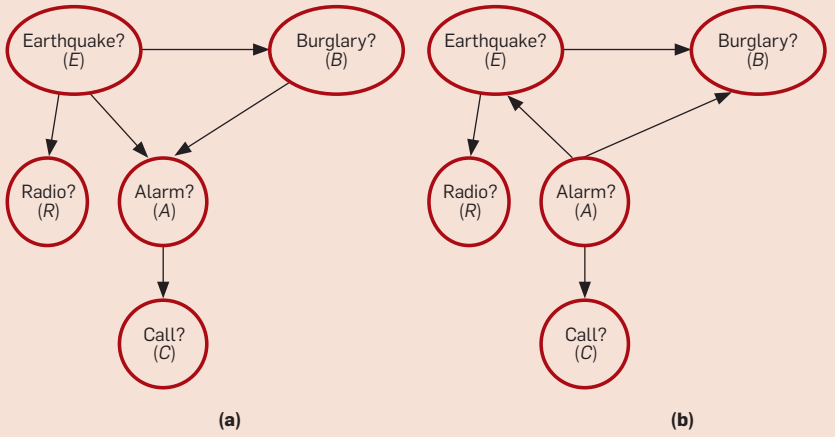
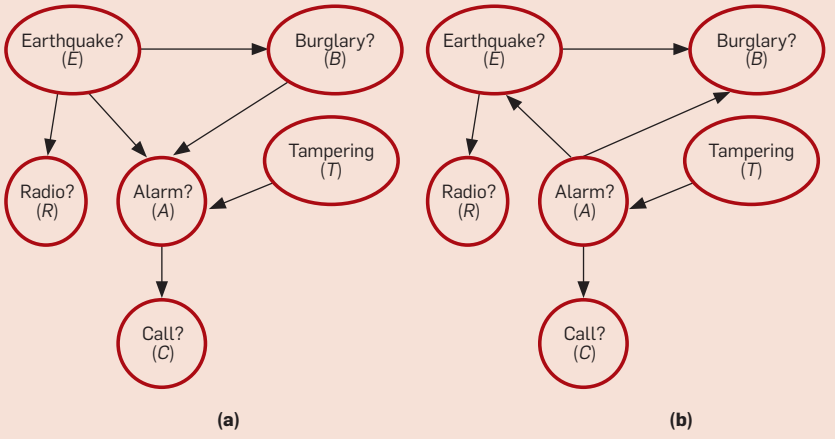


Figure 12. Extending Bayesian networks to account for interventions.



variables **E** in the Bayesian network, this query identifies the instantiation **q** of some variables $\mathbf{Q} \subseteq \overline{\mathbf{E}}$ that maximizes the probability $Pr(\mathbf{q}|\mathbf{e})$. Note the subtle difference with MPE queries: **Q** is a subset of variables $\overline{\mathbf{E}}$ instead of being all of these variables. MAP is a more difficult problem than MPE since its decision version is known to be NP^{PP}-complete, while MPE is only *NP*-complete. As an example of this query, consider Figure 5 and suppose we are interested in the most likely configuration of the two fans given that the system is unavailable. We can find this configuration using a MAP query with the evidence **e** being $S = \texttt{un\_avail}$ and $\mathbf{Q} = \{F_1, F_2\}$.

One can use these canonical queries to implement more sophisticated queries, such as the ones demanded by sensitivity analysis. This is a mode of analysis that allows one to check the robustness of conclusions drawn from Bayesian networks against perturbations in network parameters (for example, see Darwiche[8]). Sensitivity analysis can also be used for automatically revising these parameters in order to satisfy some global constraints that are imposed by experts, or derived from the formal specifications of tasks under consideration. Suppose for example that we compute the overall system reliability using the network in Figure 5 and it turns out to be 95%. Suppose we wish this reliability to be no less than 99%: $Pr(S = \texttt{avail}) \geq 99\%$. Sensitivity analysis can be used to identify components whose reliability is relevant to achieving this objective, together with the new reliabilities they must attain for this purpose. Note that component reliabilities correspond to network parameters in this example.

How Well Do Bayesian Networks Scale?

Algorithms for inference on Bayesian networks fall into two main categories: exact and approximate algorithms. Exact algorithms make no compromises on accuracy and tend to be more expensive computationally. Much emphasis was placed on exact inference in the 1980s and early 1990s, leading to two classes of algorithms based on the concepts of *elimination*[10,24,36] and *conditioning*.[6,29] In their pure form, the complexity of these algorithms is exponential only in the network *treewidth*, which is a graph-theoretic parameter that measures the resemblance of a graph to a tree structure. For example, the treewidth of trees is ≤ 1 and, hence, inference on such tree networks is quite efficient. As the network becomes more and more connected, its treewidth increases and so does the complexity of inference. For example, the network in Figure 9 has a treewidth of n assuming an underlying image with $n \times n$ pixels. This is usually too high, even for modest-size images, to make these networks accessible to treewidth-based algorithms.

The pure form of elimination and conditioning algorithms are called *structure-based* as their complexity is sensitive only to the network structure. In particular, these algorithms will consume the same computational resources when applied to two networks that share the same structure, regardless of what probabilities are used to annotate them. It has long been observed that inference algorithms can be made more efficient if they also exploit the structure exhibited by network parameters, including determinism (0 and 1 parameters) and context-specific independence (independence that follows from network parameters and is not detectable by d-separation[3]). Yet, algorithms for exploiting parametric structure have only matured in the last few years, allowing one to perform exact inference on some networks whose treewidth is quite large (see survey[8]). Networks that correspond to genetic linkage analysis (Figure 6) tend to fall in this category[12] and so do networks that are synthesized from relational models.[4]

One of the key techniques for exploiting parametric structure is based on compiling Bayesian networks into arithmetic circuits, allowing one to reduce probabilistic inference to a process of circuit propagation;[7] see Figure 10. The size of these compiled circuits is determined by both the network topology and its parameters, leading to relatively compact circuits in some situations where the parametric structure is excessive, even if the network treewidth is quite high (for example, Chavira et al.[4]). Reducing inference to circuit propagation makes it also easier to support applications that require real-time inference, as in certain diagnosis applications.[25]

Around the mid-1990s, a strong belief started forming in the inference community that the performance of exact algorithms must be exponential in treewidth—this is before parametric structure was being exploited effectively. At about the same time, methods for automatically constructing Bayesian networks started maturing to the point of yielding networks whose treewidth is too large to be handled efficiently by exact algorithms at the time. This has led to a surge of interest in approximate inference algorithms, which are generally independent of treewidth. Today, approximate inference algorithms are the only choice for networks that have a high treewidth, yet lack sufficient parametric structure—the networks used in low-level vision applications tend to have this property. An influential class of approximate inference algorithms is based on reducing the inference problem to a constrained optimization problem, with *loopy belief propagation* and its generalizations as one key example.[29,35] Loopy belief propagation is actually the common algorithm of choice today for handling networks with very high treewidth, such as the ones arising in vision or channel coding applications. Algorithms based on stochastic sampling have also been pursued for a long time and are especially important for inference in Bayesian networks that contain continuous variables.[8,15,22] Variational methods provide another important class of approximation techniques[19,22] and are key for inference on some Bayesian networks, such as the ones arising in topic models.[2]

Causality, Again

One of the most intriguing aspects of Bayesian networks is the role they play in formalizing causality. To illustrate this point, consider Figure 11, which depicts two Bayesian network structures over the same set of variables. One can verify using d-separation that these structures represent the same set of conditional independencies. As such, they are representationally equivalent as they can induce the same set of probability distributions when augmented with appropriate

CPTs. Note, however, that the network in Figure 11a is consistent with common perceptions of causal influences, yet the one in Figure 11b violates these perceptions due to edges $A \rightarrow E$ and $A \rightarrow B$. Is there any significance to this discrepancy? In other words, is there some additional information that can be extracted from one of these networks, which cannot be extracted from the other? The answer is yes according to a body of work on causal Bayesian networks, which is concerned with a key question:[16,30] how can one characterize the additional information captured by a causal Bayesian network and, hence, what queries can be answered only by Bayesian networks that have a causal interpretation?

According to this body of work, only causal networks are capable of updating probabilities based on interventions, as opposed to observations. To give an example of this difference, consider Figure 11 again and suppose that we want to compute the probabilities of various events given that someone has tampered with the alarm, causing it to go off. This is an intervention, to be contrasted with an observation, where we know the alarm went off but without knowing the reason. In a causal network, interventions are handled as shown in Figure 12a: by simply adding a new direct cause for the alarm variable. This local fix, however, cannot be applied to the non-causal network in Figure 11b. If we do, we obtain the network in Figure 12b, which asserts the following (using d-separation): if we observe the alarm did go off, then knowing it was not tampered with is irrelevant to whether a burglary or an earthquake took place. This independence, which is counterintuitive, does not hold in the causal structure and represents one example of what may go wrong when using a non-causal structure to answer questions about interventions.

Causal structures can also be used to answer more sophisticated queries, such as counterfactuals. For example, the probability of "the patient would have been alive had he not taken the drug" requires reasoning about interventions (and sometimes might even require functional information, beyond standard causal Bayesian networks[30]). Other types of queries include ones for distinguishing between direct and indirect causes and for determining the sufficiency and necessity of causation.[30] Learning causal Bayesian networks has also been treated,[16,30] although not as extensively as the learning of general Bayesian networks.

One of the most intriguing aspects of Bayesian networks is the role they play in formalizing causality.

Beyond Bayesian Networks

Viewed as graphical representations of probability distributions, Bayesian networks are only one of several other models for this purpose. In fact, in areas such as statistics (and now also in AI), Bayesian networks are studied under the broader class of *probabilistic graphical models*, which include other instances such as Markov networks and chain graphs (for example, Edwards[11] and Koller and Friedman[22]). Markov networks correspond to undirected graphs, and chain graphs have both directed and undirected edges. Both of these models can be interpreted as compact specifications of probability distributions, yet their semantics tend to be less transparent than Bayesian networks. For example, both of these models include numeric annotations, yet one cannot interpret these numbers directly as probabilities even though the whole model can be interpreted as a probability distribution. Figure 9b depicts a special case of a Markov network, known as a Markov random field (MRF), which is typically used in vision applications. Comparing this model to the Bayesian network in Figure 9a, one finds that smoothness constraints between two adjacent pixels P_i and P_j can now be represented by a single undirected edge $P_i - P_j$ instead of two directed edges and an additional node, $P_i \rightarrow C_{ij} \leftarrow P_j$. In this model, each edge is associated with a function $f(P_i, P_j)$ over the states of adjacent pixels. The values of this function can be used to capture the smoothness constraint for these pixels, yet do not admit a direct probabilistic interpretation.

Bayesian networks are meant to model probabilistic beliefs, yet the interest in such beliefs is typically motivated by the need to make rational decisions. Since such decisions are often contemplated in the presence of uncertainty, one needs to know the likelihood and utilities associated with various decision outcomes. A

classical example in this regard concerns an oil wildcatter that needs to decide whether or not to drill for oil at a specific site, with an additional decision on whether to request seismic soundings that may help determine the geological structure of the site. Each of these decisions has an associated cost. Moreover, their potential outcomes have associated utilities and probabilities. The need to integrate these probabilistic beliefs, utilities and decisions has lead to the development of *Influence Diagrams*, which are extensions of Bayesian networks that include three types of nodes: chance, utility, and decision.[18] Influence diagrams, also called decision networks, come with a toolbox that allows one to compute optimal strategies: ones that are guaranteed to produce the highest expected utility.[20,22]

Bayesian networks have also been extended in ways that are meant to facilitate their construction. In many domains, such networks tend to exhibit regular and repetitive structures, with the regularities manifesting in both CPTs and network structure. In these situations, one can synthesize large Bayesian networks automatically from compact high-level specifications. A number of concrete specifications have been proposed for this purpose. For example, template-based approaches require two components for specifying a Bayesian network: a set of network templates whose instantiation leads to network segments, and a specification of which segments to generate and how to connect them together.[22,23] Other approaches include languages based on first-order logic, allowing one to reason about situations with varying sets of objects (for example, Milch et al.[26]).

The Challenges Ahead

Bayesian networks have been established as a ubiquitous tool for modeling and reasoning under uncertainty. The reach of Bayesian networks, however, is tied to their effectiveness in representing the phenomena of interest, and the scalability of their inference algorithms. To further improve the scope and ubiquity of Bayesian networks, one therefore needs sustained progress on both fronts. The main challenges on the first front lie in increasing the expressive power of Bayesian network representations, while maintaining the key features that have proven necessary for their success: modularity of representation, transparent graphical nature, and efficiency of inference. On the algorithmic side, there is a need to better understand the theoretical and practical limits of exact inference algorithms based on the two dimensions that characterize Bayesian networks: their topology and parametric structure.

With regard to approximate inference algorithms, the main challenges seem to be in better understanding their behavior to the point where we can characterize conditions under which they are expected to yield good approximations, and provide tools for practically trading off approximation quality with computational resources. Pushing the limits of inference algorithms will immediately push the envelope with regard to learning Bayesian networks since many learning algorithms rely heavily on inference. One cannot emphasize enough the importance of this line of work, given the extent to which data is available today, and the abundance of applications that require the learning of networks from such data. C

References

1. Bayes, T. An essay towards solving a problem in the doctrine of chances. *Phil. Trans. 3* (1963), 370–418. Reproduced in W.E. Deming.
2. Blei, D.M., Ng, A.Y. and Jordan, M.I. Latent Dirichlet allocation. *Journal of Machine Learning Research 3* (2003), 993–1022.
3. Boutilier, C., Friedman, N., Goldszmidt, M. and Koller, D. Context-specific independence in Bayesian networks. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence* (1996), 115–123.
4. Chavira, M., Darwiche, A. and Jaeger, M. Compiling relational Bayesian networks for exact inference. *International Journal of Approximate Reasoning 42*, 1-2 (May 2006) 4–20.
5. Cowell, R., Dawid, A., Lauritzen, S. and Spiegelhalter, D. *Probabilistic Networks and Expert Systems.* Springer, 1999.
6. Darwiche, A. Recursive conditioning. *Artificial Intelligence 126*, 1-2 (2001), 5–41.
7. Darwiche, A. A differential approach to inference in Bayesian networks. *Journal of the ACM 50*, 3 (2003).
8. Darwiche, A. *Modeling and Reasoning with Bayesian Networks.* Cambridge University Press, 2009.
9. Dean, T. and Kanazawa, K. A model for reasoning about persistence and causation. *Computational Intelligence 5*, 3 (1989), 142–150.
10. Dechter, R. Bucket elimination: A unifying framework for probabilistic inference. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence* (1996), 211–219.
11. Edwards, D. *Introduction to Graphical Modeling.* Springer, 2nd edition, 2000.
12. Fishelson, M. and Geiger, D. Exact genetic linkage computations for general pedigrees. *Bioinformatics 18*, 1 (2002), 189–198.
13. Frey, B. editor. *Graphical Models for Machine Learning and Digital Communication.* MIT Press, Cambridge, MA, 1998.
14. Friedman, N., Geiger, D. and Goldszmidt, M. Bayesian network classifiers. *Machine Learning 29*, 2-3 (1997), 131–163.
15. Gilks, W., Richardson, S. and Spiegelhalter, D. *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics.* Chapman & Hall/CRC, 1995.
16. Glymour, C. and Cooper, G. eds. *Computation, Causation, and Discovery.* MIT Press, Cambridge, MA, 1999.
17. Heckerman, D. A tutorial on learning with Bayesian networks. *Learning in Graphical Models.* Kluwer, 1998, 301–354.
18. Howard, R.A. and Matheson, J.E. Influence diagrams. *Principles and Applications of Decision Analysis, Vol. 2.* Strategic Decision Group, Menlo Park, CA, 1984, 719–762.
19. Jaakkola, T. Tutorial on variational approximation methods. *Advanced Mean Field Methods.* D. Saad and M. Opper, ed, MIT Press, Cambridge, MA, 2001, 129–160.
20. Jensen, F.V. and Nielsen, T.D. *Bayesian Networks and Decision Graphs.* Springer, 2007.
21. Jordan, M., Ghahramani, Z., Jaakkola, T. and Saul, L. An introduction to variational methods for graphical models. *Machine Learning 37*, 2 (1999), 183–233.
22. Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, Cambridge, MA, 2009.
23. Koller, D. and Pfeffer, A. Object-oriented Bayesian networks. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence* (1997), 302–313.
24. Lauritzen, S.L. and Spiegelhalter, D.J. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of Royal Statistics Society, Series B 50*, 2 (1998), 157–224.
25. Mengshoel, O., Darwiche, A., Cascio, K., Chavira, M., Poll, S. and Uckun, S. Diagnosing faults in electrical power systems of spacecraft and aircraft. In *Proceedings of the 20th Innovative Applications of Artificial Intelligence Conference* (2008), 1699–1705.
26. Milch, B., Marthi, B., Russell, S., Sontag, D., Ong, D. and Kolobov, A. BLOG: Probabilistic models with unknown objects. In *Proceedings of the International Joint Conference on Artificial Intelligence* (2005), 1352–1359.
27. Neapolitan, R. *Learning Bayesian Networks.* Prentice Hall, Englewood, NJ, 2004.
28. Pearl, J. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the Cognitive Science Society* (1985), 329–334.
29. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.
30. Pearl, J. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, 2000.
31. Durbin, A.K.R., Eddy, S. and Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, 1998.
32. Smyth, P., Heckerman, D. and Jordan, M. Probabilistic independence networks for hidden Markov probability models. *Neural Computation 9*, 2 (1997), 227–269.
33. Steyvers, M. and Griffiths, T. Probabilistic topic models. *Handbook of Latent Semantic Analysis.* T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, eds. 2007, 427–448.
34. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M.F. and Rother, C. A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. *IEEE Trans. Pattern Anal. Mach. Intell. 30*, 6 (2008), 1068–1080.
35. Yedidia, J., Freeman, W. and Weiss, Y. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory 1*, 7 (2005), 2282–2312.
36. Zhang, N.L. and Poole, D. A simple approach to Bayesian network computations. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, (1994), 171–178.

Adnan Darwiche (darwiche@cs.ucla.edu) is a professor and former chair of the computer science department at the University of California, Los Angeles, where he also directs the Automated Reasoning Group.

research highlights

DOI:10.1145/1859204.1859228

Technical Perspective
Iterative Signal Recovery From Incomplete Samples

By Michael Elad and Raja Giryes

YOU ARE GIVEN a large set of data values, and you are requested to compress, clean, recover, recognize, and/or predict it. Sounds familiar? This is a fundamental list of scientific tasks encountered often in our daily lives. Indeed, this is the essence of the fields of signal- and image-processing, machine-learning, data-mining, and statistics in general. Common to these and many other tasks is the heavy reliance on *models*. It all starts with our expectation that the given data is not an arbitrary set of randomly chosen numbers, and there are some dependencies and guidelines the data follows, implying the true dimensionality of the data is far smaller than the embedding space dimension.

Knowing these rules enable the solution of all the tasks mentioned here, which begs the question: *How would we know these rules to begin with?* The bad news is we will probably never know these rules. The good news is we could approximate them using a model. A model is a flexible mathematical construction that is assumed to describe the data reasonably well. The better the model, the better the treatment the data gets. The progress made through the years in processing data is a reflection of the progress in the chosen models and their sophistication.

In the past decade, much effort has been devoted to the study of one specific and universal model that is based on sparsity.[1] In a nutshell, this model assumes the data can be described by a matrix (the dictionary) multiplying a sparse vector (the representation). Extensive work in the past decade provided solid theoretical foundations for this model, studied numerical schemes for using it in practice, and explored various applications where state-of-the-art results are obtained.

The following paper by Deanna Needell and Joel Tropp belongs to this field, employing this very model to a task called Compressive Sampling (CoSa). In CoSa, the aim is to "sample" the given data by multiplying it by a matrix that projects it to a (much) lower dimension. This way, few projections are expected to characterize the complete signal, giving a compression effect. The research in CoSa concentrates on the choice of the projections to apply, algorithms for recovering the data from the projections, and most important of all, deriving bounds on the required number of projections to enable a recovery of the original data.

How should the data be recovered from the projections? Owing to the sparsity of the data's representation, we should seek the signal with the sparsest representation that is close to the given projected vector. This problem is known to be *NP*-hard and the literature offers a series of "pursuit" algorithms for its approximation. One of the greatest surprises in this field is a set of theoretical guarantees for the success of those pursuit methods, essentially claiming that for a sparse enough representation, the approximated solution is not far from ideal.

The authors propose a greedy iterative pursuit algorithm called CoSaMP, and provides a theoretical analysis of its performance, giving such a guarantee for its successful operation. This work is unique in several important ways:

▸ In the realm of pursuit methods, there are simple and crude greedy methods, and there are more complex and more accurate relaxation techniques. CoSaMP enjoys both worlds—it has the elegance and simplicity of the greedy methods, while its guarantees are similar to those encountered by the relaxation techniques. As opposed to plain greedy algorithms that accumulate the non-zeros in the representation sequentially, CoSaMP is also capable of punning non-zero entries in the representation based on their weak contribution.

The authors' theoretical analysis introduces a brilliant and powerful language, adequate for an analysis of general greedy methods.

▸ This work does much more than proposing an algorithm. Their accompanying theoretical analysis introduces a new, brilliant and powerful language, adequate for an analysis of general greedy methods. As such, it has already been used in follow-up works.[2,3]

▸ From a practical standpoint, the impact of the CoSaMP algorithm goes well beyond CoSa, despite the name chosen. CoSaMP is a general pursuit method, and as such, it is relevant to many other applications, such as denoising and general inverse problems.

▸ From a theoretical point of view, there are two important extensions: The given study is shadowed by the assumption that the representation basis should be orthonormal, whereas in many cases one would be interested in frames. Secondly, the noise is assumed to be adversarial, and as such it necessarily leads to weak guarantees. Replacing this assumption by random noise can be shown to give near-oracle performance.[3]

To conclude, CoSa, and sparse approximation in general, pose an appealing model and ways to use it successfully for various tasks. The following paper on CoSaMP is an important milestone in our journey to better understand the potential and implications of this research arena.

References

1. Bruckstein, A.M., Donoho, D.L., and Elad, M. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review 51*, 1 (Feb. 2009), 34–81.
2. Dai, W. and Milenkovic, O. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Info. Theory 55*, 5 (2009).
3. Giryes, R. and Elad, M. RIP-based near-oracle performance guarantees for subspace-pursuit, CoSaMP, and iterative hard-thresholding. Submitted to *IEEE Trans. Info. Theory*.

Michael Elad is a professor and **Raja Giryes** is a Ph.D. candidate in the computer science department at the Technion—Israel Institute of Technology, Haifa.

DOI:10.1145/1859204.1859229

CoSaMP: Iterative Signal Recovery from Incomplete and Inaccurate Samples

By Deanna Needell and Joel A. Tropp

Abstract

Compressive sampling (CoSa) is a new paradigm for developing data sampling technologies. It is based on the principle that many types of vector-space data are *compressible*, which is a term of art in mathematical signal processing. The key ideas are that randomized dimension reduction preserves the information in a compressible signal and that it is possible to develop hardware devices that implement this dimension reduction efficiently. The main computational challenge in CoSa is to reconstruct a compressible signal from the reduced representation acquired by the sampling device. This extended abstract describes a recent algorithm, called `CoSaMP`, that accomplishes the data recovery task. It was the first known method to offer near-optimal guarantees on resource usage.

1. WHAT IS COMPRESSIVE SAMPLING?

In many applications, the ambient dimension of a data vector does not reflect the actual number of degrees of freedom in that vector. In mathematical signal processing, this property is captured by the notion of a *compressible signal*. Natural images provide a concrete example of compressible signals because we can approximate them accurately just by summarizing the solid areas (local averages) and the edges (local differences). This representation allows us to compress images dramatically without a substantial loss of visual quality—an idea implemented in the JPEG image coders that we use every day.

Sampling is a generic term for the process of acquiring data about an object, either in the physical or the virtual world. There are many different technologies that we use to collect samples. A ubiquitous piece of sampling hardware is the CCD array inside a digital camera, which measures the average light intensity at each small square in the focal plane of the camera. Other sampling equipment includes X-ray machines, magnetic resonance imaging (MRI), and analog-to-digital converters.

In most cases, we use sampling to acquire as much information as possible about an object of interest. Indeed, when purchasing a digital camera, we often think that more pixels are better. But as soon as the image is captured, the camera invokes compression algorithms to reduce the amount of information it needs to store. This fact suggests an awkward question: if there is a limited amount of information in the object, why are we taking so many samples? Is there some way to obtain measurements that automatically sieve out the information in the object?

One way to exploit the gap between nominal dimension and degrees of freedom is to perform *dimension reduction*. This idea has a long tradition in computer science. Indeed, to solve geometric problems involving high-dimensional data, we often map the data to a lower dimension while approximately preserving the geometric properties that are relevant to the computation. One standard method for achieving this goal is to use a random projection to the data.

The main idea in *compressive sampling* (CoSa) is to develop sampling technologies based on dimension reduction. The mathematical foundation for this approach is the fact that an appropriate random projection of a compressible signal retains most of the information in that signal. The coordinates of this randomly projected signal are interpreted as samples, and the number of samples we need is comparable with the information content of the object we are sampling. Because we have a lot of flexibility when designing projections, we have flexibility in engineering the sampling hardware.

An intriguing example of a CoSa sensing device is the Rice University single-pixel camera. This camera uses a digital micromirror device (DMD) to selectively black out a random subset of pixels from an image and to measure the total intensity of the remaining pixels. By applying many random masks in sequence, we obtain a collection of samples that captures the information in the image.

Another key fact about CoSa is that we cannot use simple linear techniques to reconstruct the data vector from the random samples. Instead, we must develop more sophisticated algorithms. One approach that has received a lot of attention is convex programming based on ℓ_1 minimization. This abstract describes an algorithm, called `CoSaMP`, that is based on the greedy pursuit schema.[18, 19] `CoSaMP` was the first computational technique that provably achieved near-optimal guarantees on resource usage. Using the techniques from our paper, researchers later demonstrated that other algorithms offer comparable performance guarantees. Another algorithm similar to `CoSaMP`, called Subspace Pursuit,[9] was published at the same time, but the accompanying analysis was less complete.

The original version of this paper appeared in *Applied and Computational Harmonic Analysis 26*, 3 (2008), 301–321.

The remainder of the article is organized as follows. Section 2 describes the mathematical framework for CoSa as well as the theoretical results for the CoSaMP algorithm. Section 3 presents a description of the algorithm and its implementation. Section 4 discusses related work and a comparison to our method. We conclude in Section 5 with the analysis of the algorithm and sketch the proof of the main result.

2. THE MATHEMATICS OF CoSa

We continue with a description of the mathematical model for signals, sampling technologies, and signal reconstruction algorithms. The main theoretical results for the CoSaMP algorithm appear at the end of the section.

2.1. Notation

Let us instate some notation that is used throughout the paper. A *signal* is a vector $\boldsymbol{x} \in \mathbb{C}^N$. For $p \in [1, \infty]$, we write $\|\cdot\|_p$ for the usual ℓ_p vector norm: $\|\boldsymbol{x}\|_p = (\sum x_i^p)^{1/p}$. We reserve $\|\cdot\|$ for the spectral norm. For $\boldsymbol{x} \in \mathbb{C}^N$, we write $\boldsymbol{x}_r$ for the signal in $\mathbb{C}^N$ that is formed by restricting $\boldsymbol{x}$ to its r largest-magnitude components (with ties broken lexicographically).

For $T \subset \{1, 2, \ldots, N\}$, we denote the restriction of the signal to the set T as $\boldsymbol{x}|_T$, which is the vector equal to $\boldsymbol{x}$ on T and 0 elsewhere. We occasionally abuse the notation and treat $\boldsymbol{x}|_T$ as an element of $\mathbb{C}^T$. We also define the restriction $\boldsymbol{\Phi}_T$ of the sampling matrix $\boldsymbol{\Phi}$ to be the column submatrix whose columns are listed in the index set T. Finally, we define the pseudoinverse of a tall, full-rank matrix $\boldsymbol{A}$ by the formula $\boldsymbol{A}^\dagger = (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*$.

2.2. Sparse and compressible signals

We say that a signal is *s-sparse* when it has $s \ll N$ nonzero entries:

$$\|\boldsymbol{x}\|_0 := |\operatorname{supp}(\boldsymbol{x})| := |\{j : x_j \neq 0\}| \leq s.$$

This definition encapsulates the idea that the signal contains far less information than its ambient dimension suggests. Observe that it takes roughly $\log_2\binom{N}{s} = \mathrm{O}(s \log(N/s))$ bits to write down the locations of the nonzero entries in an s-sparse signal in $\mathbb{C}^N$.

While sparse signals are important for theory, they rarely arise in practice, so we consider the larger class of *compressible signals*. For $p \in (0, 1)$, we say that $\boldsymbol{x}$ is *p-compressible* with magnitude R if its components taken in sorted order obey

$$|x|_{(i)} \leq R \cdot i^{-1/p} \qquad \text{for } i = 1, 2, 3, \ldots \tag{1}$$

Compressible signals are well approximated by sparse signals. The smaller the number p, the better the approximation:

$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{x}_s\|_1 &\leq \mathrm{C}_p \cdot R \cdot s^{1-1/p} \\ \|\boldsymbol{x} - \boldsymbol{x}_s\|_2 &\leq \mathrm{D}_p \cdot R \cdot s^{1/2-1/p}. \end{aligned} \tag{2}$$

Let us emphasize that we do not know in advance *which* coordinates of a compressible signal are significant—only that few coordinates matter.

We limit our attention to signals that are sparse or compressible in the standard coordinate basis, although the ideas apply equally well to signals that are sparse or compressible with respect to other orthonormal bases. The generalization is important in applications. For instance, natural images are usually not compressible in the standard basis but they are compressible in an appropriate wavelet basis.

2.3. Modeling the sampling operation

We can model each of the sampling technologies mentioned in the introduction as a linear map from a signal of interest to a collection of samples. We use an $m \times N$ matrix $\boldsymbol{\Phi}$ to describe the sampling operation. Thus, we write $\boldsymbol{u} = \boldsymbol{\Phi}\boldsymbol{x}$ to summarize the process of collecting a vector $\boldsymbol{u}$ of m samples from an N-dimensional signal $\boldsymbol{x}$.

Restricted Isometry Constants: In this section, to build intuition, we focus on the class of s-sparse signals. It is important to understand what properties of the sampling matrix $\boldsymbol{\Phi}$ ensure that s-sparse signals can be distinguished based on their samples. In other words, we need to make sure that $\boldsymbol{\Phi}$ is a linear embedding (i.e., a bijective structure-preserving map) of the set of s-sparse signals into $\mathbb{C}^m$. A necessary and sufficient condition is that each $2s$-column submatrix of the sampling matrix $\boldsymbol{\Phi}$ forms a linearly independent set. Certain Vandermonde matrices with $m = 2s$ rows, arising from Reed–Solomon codes, have this property. Unfortunately, these matrices contain nearly singular $2s$-column submatrices, so the sampling matrix is not stably invertible on the image of the set of s-sparse signals. Noise and signal perturbations become a serious problem.

To avoid this difficulty, Candès and Tao[4] insist that each $2s$-column submatrix of the measurement matrix should be well conditioned. More formally, they define the rth *restricted isometry constant* of the matrix $\boldsymbol{\Phi}$ to be the least number δ_r such that

$$(1 - \delta_r)\|\boldsymbol{x}\|_2^2 \leq \|\Phi\boldsymbol{x}\|_2^2 \leq (1 + \delta_r)\|\boldsymbol{x}\|_2^2 \text{ for } \|\boldsymbol{x}\|_0 \leq r. \tag{3}$$

The condition $\delta_{2s} \ll 1$ implies that the measurement matrix $\boldsymbol{\Phi}$ preserves the geometry (with respect to the Euclidean metric) of the set of all s-sparse signals. This condition is sufficient to ensure that sampling is stably invertible. It has become standard practice in CoSa to assume that the measurement matrix satisfies a condition of this type.

How Many Samples?: The next question is *how many* samples m do we need to ensure that the sampling map embeds the set of s-sparse signals into $\mathbb{C}^m$. Because an s-sparse signal contains only about $s \log(N/s)$ bits of information, we hope that $m = \mathrm{O}(s \log(N/s))$ samples suffice for the embedding. In fact, many random sampling schemes allow us to achieve this sampling rate, or one that is just slightly worse. (It is *necessary* in some sense to take at least this many samples, so these sampling schemes are optimal.) Furthermore, there are technologies that implement these sampling schemes. Here are two examples.

A *random sign matrix* has independent, identically distributed entries that take the values ± 1 with equal probability. Its restricted isometry constant[5] satisfies

$$\delta_{2s} \leq \varepsilon \quad \text{whenever} \quad m \gtrsim \frac{s \log(N/s)}{\varepsilon^2}.$$

In words, we can embed the set of s-sparse signals into m dimensions, where m is proportional to the sparsity and

logarithmic in the ambient dimension. A random sign matrix can be used to construct the random masks required by the Rice single-pixel camera.

The *subsampled Fourier matrix* consists of a collection of m rows chosen uniformly at random from the discrete Fourier transform matrix. The best theoretical bound for the restricted isometry constant[23] is

$$\delta_{2s} \le \varepsilon \quad \text{whenever} \quad m \gtrsim \frac{s \log^5(N) \cdot \log(\varepsilon^{-1})}{\varepsilon^2}$$

It is conjectured that the actual scaling is $m \gtrsim s \log(N)$. The subsampled Fourier matrix describes measurements that can be acquired by an MRI machine. Note that a subsampled Fourier matrix can be applied to a vector using $O(N \log N)$ arithmetic operations, and it requires only $O(m \log N)$ bits of storage. These computational properties are very important in practice, as we will see.

Certain types of deterministic matrices also satisfy bounds on their restricted isometry constants. At present, all known examples require that the number m of samples satisfy $m \gtrsim s^2$. Randomness seems to provide a significant advantage when designing schemes for CoSa.

2.4. Signal reconstruction via CoSaMP

When the sampling matrix $\mathbf{\Phi}$ stably embeds the set of s-sparse signals, it is possible in principle to recover an arbitrary s-sparse signal from its samples. The major computational question in CoSa is to determine how to reconstruct the sparse signal *using a tractable algorithm*. For practical applications, we also need to understand how the reconstruction is affected when the signal is merely compressible and the samples are contaminated by noise. The following issues are crucial:

1. Can the algorithm recover the signal under a variety of sampling schemes? In other words, is the class of measurement maps for which the algorithm can succeed large?
2. Can the algorithm reconstruct the signal from a minimal number m of samples?
3. Is signal recovery robust when samples are contaminated with noise or the signal is not exactly sparse? Is the error bound optimal for each target signal?
4. Does the algorithm offer provably efficient computational cost and storage?

This abstract discusses a reconstruction algorithm called Compressive Sampling Matching Pursuit (CoSaMP) that satisfies all of the foregoing criteria. Our original publications[18,19] were the first to prove that a CoSa signal reconstruction algorithm could achieve essentially optimal resource usage.

Properties of the CoSaMP Algorithm: Let us state and explain our main result. We postpone the details of the algorithm and the analysis until later.

THEOREM A (CoSaMP). *Suppose that $\mathbf{\Phi}$ is an $m \times N$ sampling matrix with restricted isometry constant $\delta_{2s} \le c$. Let $\boldsymbol{u} = \mathbf{\Phi}\boldsymbol{x} + \boldsymbol{e}$ be a vector of samples of an arbitrary signal, contaminated with arbitrary noise.*

The CoSaMP *algorithm takes as input a routine to multiply $\mathbf{\Phi}$ and $\mathbf{\Phi}^*$ by a vector, the vector $\boldsymbol{u}$ of samples, the sparsity parameter s, and a precision parameter η. The output is an s-sparse signal approximation $\boldsymbol{a}$ that satisfies*

$$\|\boldsymbol{x} - \boldsymbol{a}\|_2 \le C \cdot \max\left\{\eta, \frac{1}{\sqrt{s}}\|\boldsymbol{x} - \boldsymbol{x}_{s/2}\|_1 + \|\boldsymbol{e}\|_2\right\}.$$

The running time is $O(\mathcal{L} \cdot \log(\|\boldsymbol{x}\|_2/\eta))$, where $\mathcal{L}$ bounds the cost of a matrix–vector multiply with $\mathbf{\Phi}$ or $\mathbf{\Phi}^*$. Working storage is $O(N)$. The constants $c < 1$ and $C > 1$ are absolute.

How does this result deliver the desired performance?

1. The algorithm requires only that the matrix satisfy a bound on the restricted isometry constant. Many types of random matrices (and certain deterministic matrices) have this property. As a consequence, the algorithm can be used with many measurement technologies.
2. The number m of samples required to obtain the bound $\delta_{2s} \le c$ depends on the type of sampling matrix. For many types of random matrices, $m = O(s \log^{\alpha} N)$ for a small integer α. Since this value is optimal up to the constants, the algorithm can recover signals from a minimal number of samples.
3. The error bound demonstrates that the algorithm succeeds for all signals, even when there is noise in the samples. As we expect, the algorithm performs best for sparse and compressible signals. Indeed, Equation 2 shows that CoSaMP produces an approximation $\boldsymbol{a}$ of a p-compressible signal $\boldsymbol{x}$ that satisfies

$$\|\boldsymbol{x} - \boldsymbol{a}\|_2 \le C \cdot \max\left\{\eta, R \cdot s^{1/2 - 1/p} + \|\boldsymbol{e}\|_2\right\},$$

   which is the best possible approximation error we could hope for in general.[6]
4. The computational cost of the algorithm depends on the type of signals we are approximating. For a compressible signal, we may take the precision parameter $\eta = R \cdot s^{1/2 - 1/p}$ to see that the number of iterations required to produce a minimal approximation error is at most $O(\log s)$. See Section 1.2 of Needell and Tropp[19] for details.

   The running time also depends on the type of sampling matrix we use. In particular, we can apply a subsampled Fourier matrix to a vector using $\mathcal{L} = O(N \log N)$ operations.

   It follows that if we acquire a compressible signal with a structured sampling matrix, we can perform signal recovery in time $O(N \log^2 N)$. This runtime bound is essentially optimal when the sparsity level is roughly the same order as the dimension (which is the setting in most practical problems).

3. THE CoSaMP ALGORITHM

The CoSaMP algorithm is, at heart, a greedy iterative method for reconstructing a signal from compressive samples. This section provides an overview of the algorithm and its implementation. We also explain the basic idea behind the analysis of the algorithm, which demonstrates that each iteration

reduces the error in the current signal approximation.

3.1. Description of algorithm

The input to the algorithm consists of (access to) the sampling operator $\mathbf{\Phi}$, the samples $\boldsymbol{u}$, the target sparsity level s, and a halting criterion. We impose the following hypotheses:

- The sparsity level s is fixed, and the $m \times N$ sampling operator $\mathbf{\Phi}$ has restricted isometry constant $\delta_{4s} \leq 0.1$.
- The signal $\boldsymbol{x} \in \mathbb{C}^N$ is arbitrary, except where noted. The noise vector $\boldsymbol{e} \in \mathbb{C}^m$ is also arbitrary.
- The vector of samples $\boldsymbol{u} = \mathbf{\Phi}\boldsymbol{x} + \boldsymbol{e}$.

The algorithm is initialized with a trivial signal estimate, meaning that the initial residual is the entire unknown target signal. Each iteration then consists of five major steps.

1. **Identification.** Using the current samples, the algorithm computes a vector that is highly correlated with the signal, called the signal proxy. From the proxy, components of the signal that carrry a lot of energy are located.
2. **Support Merger.** The set of newly identified components is united with the set of components that appear in the current approximation.
3. **Estimation.** The algorithm solves a least-squares problem to approximate the target signal on the merged set of components.
4. **Pruning.** The algorithm produces a new approximation by retaining only the largest entries in this least-squares signal approximation.
5. **Sample Update.** Finally, the samples are updated so that they reflect the residual, the part of the signal that has not been approximated.

These five steps are repeated until the halting criterion is satisfied. In our analysis, we assume that the method uses a fixed number of iterations, but Needell and Tropp[18,19] discuss other alternatives. The CoSaMP algorithm can be summarized by the pseudocode presented as Algorithm 1.

REMARK 1. *We emphasize that the method we present is a specific example from a more general* framework. *The articles[18, 19] discuss a number of variations. In particular, note that the term* $2s$ *in the identification step can be replaced by* αs *for other values of* α. *This type of tuning is valuable in practice, but it makes the proof more complicated*.

REMARK 2. *Some knowledge about the sparsity level is required as input to the algorithm. There are several strategies for estimating this parameter if it is not known a priori. One such method would be to use the number m of measurements to deduce the sparsity level. Since* $m \approx 2s \log N$ *is often sufficient, the estimate* $s \approx m/2 \log N$ *is reasonable. A second approach would be to run the* CoSaMP *algorithm using a variety of sparsity levels and select the best approximation* $\boldsymbol{a}$ *obtained using the empirical error* $\|\mathbf{\Phi}\boldsymbol{a} - \boldsymbol{u}\|_2$. *If one varies s according to a geometric progression (e.g.,* $s = 1, 2, 4, 8, \ldots, m$*), this variation increases the runtime by at most* $O(\log m)$. *See Appendix A of Needell and Tropp[19] for more details*.

Algorithm 1: CoSaMP Recovery Algorithm

Input: Sampling matrix $\mathbf{\Phi}$, noisy sample vector $\boldsymbol{u}$, sparsity level s

Output: An s-sparse approximation $\boldsymbol{a}$ of the target signal

$\boldsymbol{a}^0 \leftarrow \mathbf{0}, \boldsymbol{v} \leftarrow \boldsymbol{u}, k \leftarrow 0$ { Initialization }

repeat

$k \leftarrow k + 1$

$\boldsymbol{y} \leftarrow \mathbf{\Phi}^*\boldsymbol{v}$ { Form signal proxy }

$\Omega \leftarrow \text{supp}(\boldsymbol{y}_{2s})$ { Identify large components }

$T \leftarrow \Omega \cup \text{supp}(\boldsymbol{a}^{k-1})$ { Merge supports }

$\boldsymbol{b}|_T \leftarrow \mathbf{\Phi}_T^\dagger u$ { Signal estimation }

$\boldsymbol{b}|_{T^c} \leftarrow \mathbf{0}$

$\boldsymbol{a}^k \leftarrow \boldsymbol{b}_s$ { Prune approximation }

$\boldsymbol{v} \leftarrow \boldsymbol{u} - \mathbf{\Phi}\boldsymbol{a}^k$ { Update current samples }

until halting criterion *true*

3.2. Implementation

CoSaMP was designed to be a practical algorithm for signal recovery, so it is worth spending some time on implementation issues. The least-squares step in the algorithm presents the largest contribution to the runtime. Fortunately, we have constructed $\mathbf{\Phi}_T$ to ensure it is well conditioned, so we can apply its pseudoinverse quickly using an iterative method. Section 5 of Needell and Tropp[19] contains a thorough analysis of iterative least-squares methods as modules in CoSaMP. This analysis shows that the cost of solving the least-squares problem is $O(\mathcal{L})$, where $\mathcal{L}$ bounds the cost of a matrix–vector multiply with $\mathbf{\Phi}_T$ or $\mathbf{\Phi}_T^*$.

The remaining steps of the algorithm involve standard techniques, and we can estimate the operation counts as follows.

Proxy. Forming the proxy is dominated by the cost of the matrix–vector multiply $\mathbf{\Phi}^*\boldsymbol{v}$.

Identification. We can locate the largest $2s$ entries of a vector in time $O(N)$ using the approach in Cormen et al.[7, Ch. 9]. In practice, it may be faster to use an $O(N \log N)$ sorting algorithm (e.g., quicksort, mergesort, heapsort[7, Sec. II]) on the entries of the signal and then select the first $2s$ of them.

Support Merger. We can merge two support sets of size $O(s)$ in expected time $O(s)$ via randomized hashing methods,[7, Ch. 11] or by sorting both sets first and using the elementary merge procedure[7, p. 29] for a total cost $O(s \log s)$.

LS estimation. We can use the conjugate gradient method or the LSQR algorit (see e.g. Paige and Saunders[22]) to compute $\mathbf{\Phi}_T^\dagger \boldsymbol{u}$. Initializing the least-squares algorithm requires a matrix–vector multiply with $\mathbf{\Phi}_T^*$, and each iteration requires one matrix–vector multiply each with $\mathbf{\Phi}_T$ and $\mathbf{\Phi}_T^*$. Since $\mathbf{\Phi}_T$ is a submatrix of $\mathbf{\Phi}$, the matrix–vector multiplies can also be obtained from multiplication with the full matrix. We show in Section 5 of Needell and Tropp[19] that a constant number of least-squares iterations suffices for our results to hold.

Pruning. As in the identification step, pruning can be

Table 1. Operation count for CoSaMP

Step	Standard Multiply	Fast Multiply
Form proxy	mN	$\mathcal{L}$
Identification	N	N
Support merger	s	s
LS estimation	sm	$\mathcal{L}$
Pruning	s	s
Sample update	sm	$\mathcal{L}$
Total per iteration	$O(mN)$	$O(\mathcal{L})$

implemented in time $O(s)$, but it may be preferable to sort the components of the vector and then select the first s at a cost of $O(s \log s)$.

Sample Update. This step is dominated by the cost of the multiplication of $\mathbf{\Phi}$ with the s-sparse vector $\boldsymbol{a}^k$.

Table 1 summarizes the cost of each step for the cases in which the standard and fast multiply is used.

Finally, we may analyze the storage requirements of the algorithm. Aside from the sampling matrix storage requirement, the algorithm constructs only one vector of length N, the signal proxy. The sample vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ have length m, and thus require $O(m)$ storage. Since the signal approximations are sparse, they can be stored using sparse data structures which require at most $O(s \log N)$ storage. Similarly, the index sets that appear require only $O(s \log N)$ storage. The total working storage is therefore $O(N)$.

The following result summarizes these observations.

THEOREM 1 (RESOURCE REQUIREMENTS). *Each iteration of* CoSaMP *requires* $O(\mathcal{L})$ *time, where* $\mathcal{L}$ *bounds the cost of a multiplication with the matrix* $\mathbf{\Phi}$ *or* $\mathbf{\Phi}^*$. *The algorithm uses working storage* $O(N)$.

3.3. Error reduction bound

The theoretical analysis of the CoSaMP algorithm is based on an estimate for how much the algorithm reduces the approximation error in each iteration. This result demonstrates that the algorithm makes significant progress whenever the error is substantially larger than a certain baseline value. We call this baseline the *unrecoverable energy* ν in the signal:

$$\nu \stackrel{\text{def}}{=} \|\boldsymbol{x} - \boldsymbol{x}_s\|_2 + \frac{1}{\sqrt{s}}\|\boldsymbol{x} - \boldsymbol{x}_s\|_1 + \|\boldsymbol{e}\|_2. \tag{4}$$

The unrecoverable energy is due to the noise in the samples and the fact that the signal is not exactly sparse. For a detailed discussion, see Section 2.5 of Needell and Tropp.[19]

The main technical result is the following iteration invariant, which we establish in Section 5.

THEOREM 2 (ERROR REDUCTION). *For each iteration* $k \geq 0$, *the signal approximation* $\boldsymbol{a}^k$ *is* s*-sparse and*

$$\|\boldsymbol{x} - \boldsymbol{a}^{k+1}\|_2 \leq 0.5\,\|\boldsymbol{x} - \boldsymbol{a}^k\|_2 + 10\nu.$$

In particular,

$$\|\boldsymbol{x} - \boldsymbol{a}^k\|_2 \leq 2^{-k}\,\|\boldsymbol{x}\|_2 + 20\nu.$$

The main consequence of Theorem 2 is the fact that, after $\log_2(\|\boldsymbol{x}\|_2/\eta)$ iterations, the approximation error is no greater than $\eta + 20\nu$. See Needell and Tropp[18, 19] for more discussion.

The statement of Theorem A depends on a simple upper bound for the ℓ_2 term in the unrecoverable energy:

$$\nu \leq \frac{1.71}{\sqrt{s}}\|\boldsymbol{x} - \boldsymbol{x}_{s/2}\|_1 + \|\boldsymbol{e}\|_2. \tag{5}$$

This estimate is a consequence of Lemma 7 of Gilbert.[15]

Finally, we can replace the assumption that $\delta_{4s} \leq 0.1$ by a stronger bound on δ_{2s} because of Corollary 3.4 in Needell and Tropp,[19] which states that $\delta_{cr} \leq c \cdot \delta_{2r}$ for any positive integers c and r.

4. RELATED WORK

CoSaMP is inspired by algorithmic ideas and analytic techniques that have appeared previously. We briefly summarize the major algorithmic approaches and compare them with CoSaMP. This work can be classified in three rough categories: convex relaxation,[4, 11] greedy pursuits,[12, 20, 24] and combinatorial algorithms.[8, 14–16]

The convex optimization methods[1, 10] attempt to reconstruct signals by solving the mathematical program

$$\min \|\boldsymbol{y}\|_1 \quad \text{subject to} \quad \|\mathbf{\Phi}\boldsymbol{y} - \boldsymbol{u}\|_2 \leq \varepsilon, \tag{6}$$

where we assume that $\|\boldsymbol{e}\|_2 \leq \varepsilon$. The intuition behind minimizing the ℓ_1 norm is that this norm promotes sparsity, and so the solution to this program should approximate a compressible signal well. Candès et al.[3] establish that each minimizer $\boldsymbol{a}$ of Equation 6 satisfies

$$\|\boldsymbol{x} - \boldsymbol{a}\|_2 \leq \mathrm{C}\left[\frac{1}{\sqrt{s}}\|\boldsymbol{x} - \boldsymbol{x}_s\|_1 + \varepsilon\right] \tag{7}$$

whenever $\mathbf{\Phi}$ has restricted isometry constant $\delta_{4s} \leq 0.2$. These restricted isometry constant requirements continue to be improved.[2, 13] The error bound for CoSaMP is equivalent with Equation 7, up to the precise value of the constants.

Many algorithms have been proposed to optimize Equation 6. In particular, interior-point methods[1, 17] are guaranteed to solve the problem to a fixed precision in $O(m^2N^{1.5})$ time.[21] The runtime for CoSaMP is much better than these interior-point methods.

Greedy algorithms such as OMP,[24] StOMP,[12] and ROMP[20] are iterative methods for approximating a signal from compressive samples. In each iteration, they use a greedy rule to identify new elements in the support of the signal. These methods provide fast runtimes. On the other hand, to the extent that their behavior is understood, greedy algorithms require more measurements or produce worse errors than the convex optimization Equation 6.

Tropp and Gilbert proposed the greedy algorithm *orthogonal matching pursuit* (OMP) for signal recovery.[24] OMP is similar to CoSaMP in that it uses a signal proxy to select large components of the target signal, but it selects one such component at each iteration. However, it does not provide the same uniform guarantees as convex relaxation, and it is unknown whether it succeeds in the noisy setting.

Other algorithms based on OMP have been formulated, such as *regularized OMP*, or ROMP,[20] developed by Needell and Vershynin. ROMP is noteworthy because the authors establish that under the restricted isometry property, the algorithm can approximately recover compressible signals from noisy samples. The error bound and measurement requirements provided by these results are analogous with those of the convex optimization method, modulo parasitic logarithmic terms. Our results for CoSaMP remove all the extraneous factors, so the performance of CoSaMP is essentially optimal.

Finally, we mention that there is a class of algorithms for signal recovery that have sublinear (in the signal dimension) runtime. Some examples are the Fourier sampling algorithm (FS) of Gilbert et al.,[16] chaining pursuit,[14] and HHS pursuit.[15] These methods are very fast, but they require highly structured measurements.

Table 2 summarizes the behavior of the above algorithms in terms of the following five criteria.

General samples. Does the algorithm succeed for a variety of sampling schemes, or does it require structured samples? The designation "RIP" implies that a bound on a restricted isometry constant suffices. "Gauss" means that the algorithm succeeds when the sampling matrix $\mathbf{\Phi}$ has iid subgaussian entries.

Optimal number of samples. Can the algorithm reconstruct s-sparse signals from a near-optimal number of measurements, $m = O(s \log N)$? Or are its sampling requirements higher (by logarithmic factors)?

Uniformity. Given a fixed sampling matrix, does the algorithm recover all signals? Or do the results hold with high probability only for each fixed signal?

Stability. Does the algorithm succeed when the signal is compressible (but not sparse) and the samples are noisy?

Running time. In the worst case (without fast multiplies), how long does it take the algorithm to recover a real-valued s-sparse signal within a fixed relative precision, given a sampling matrix with no special structure? The designation LP(N, m) represents the cost of solving a linear program with N variables and m constraints ($O(m^2N^{1.5})$ using an interior-point method).

Under all of these metrics, CoSaMP achieves the best performance out of the linear and superlinear methods. Of course, CoSaMP is slower than the sublinear algorithms, but it allows more general sampling matrices and demands fewer samples, which makes it more adaptable to practical applications. CoSaMP delivers the same guarantees as the best convex optimization approaches as well as rigorous bounds on computational cost and storage that are comparable with faster greedy methods. Thus, CoSaMP is essentially optimal in every important respect.

Table 2. Comparison of several signal recovery algorithms

	CoSaMP	ROMP	Convex Optimization
General samples	RIP	RIP	RIP
Opt. # samples	Yes	No	Yes
Uniformity	Yes	Yes	Yes
Stability	Yes	Yes	Yes
Running time	$O(mN)$	$O(smN)$	LP(N, m)

	OMP	Fourier Sampling	HHS Pursuit
General samples	Gauss	No	No
Opt. # samples	Yes	No	No
Uniformity	No	No	Yes
Stability	?	Yes	Yes
Running time	$O(smN)$	s polylog(N)	poly(s log N)

5. PROOF OF RESULTS

The CoSaMP algorithm uses an approach inspired by the restricted isometry property to identify the largest s components of the target signal. Owing to the restricted isometry property, each set of s components of the proxy vector $\boldsymbol{y} = \mathbf{\Phi}^*\mathbf{\Phi}\boldsymbol{x}$ approximates the energy in the corresponding s components of $\boldsymbol{x}$. Since the samples are of the form $\boldsymbol{u} = \mathbf{\Phi}\boldsymbol{x}$, we can obtain the proxy by applying the matrix $\mathbf{\Phi}^*$ to the samples $\boldsymbol{u}$. Once the set T of significant locations is identified, the signal coefficients can be recovered using $\mathbf{\Phi}_T^\dagger$.

The algorithm repeats this idea at each iteration, updating the samples to reflect the residual (the part of the signal yet to be approximated). We use least squares to estimate the signal on this support set and repeat this process until the recoverable energy in the signal has been found.

We now outline the proof of Theorem 2.

5.1. Consequences of the RIP

We begin with some important consequences of the restricted isometry property. Omitted proofs appear in Needell and Tropp.[19]

PROPOSITION 1 (CONSEQUENCES). *Suppose $\mathbf{\Phi}$ has restricted isometry constant δ_r. Let T be a set of r indices or fewer. Then*

$$\|\mathbf{\Phi}_T^* \boldsymbol{u}\|_2 \le \sqrt{1+\delta_r}\,\|\boldsymbol{u}\|_2$$

$$\|\mathbf{\Phi}_T^\dagger \boldsymbol{u}\|_2 \le \frac{1}{\sqrt{1-\delta_r}}\|\boldsymbol{u}\|_2$$

$$\|\mathbf{\Phi}_T^* \mathbf{\Phi}_T \boldsymbol{x}\|_2 \lessgtr (1 \pm \delta_r)\|\boldsymbol{x}\|_2$$

$$\|(\mathbf{\Phi}_T^* \mathbf{\Phi}_T)^{-1} \boldsymbol{x}\|_2 \lessgtr \frac{1}{1+\delta_r}\|\boldsymbol{x}\|_2.$$

where the last two statements contain an upper and lower bound, depending on the sign chosen.

A corollary of these bounds is the fact that disjoint sets of columns from the sampling matrix span nearly orthogonal subspaces.

PROPOSITION 2 (ORTHOGONALITY). *Suppose $\mathbf{\Phi}$ has restricted isometry constant δ_r. Let S and T be disjoint sets of indices whose combined cardinality does not exceed r. Then*

$$\|\mathbf{\Phi}_S^* \mathbf{\Phi}_T\| \le \delta_r.$$

We apply this result through a corollary.

COROLLARY 1. *Suppose $\mathbf{\Phi}$ has restricted isometry constant δ_r. Let T be a set of indices, and let $\boldsymbol{x}$ be a vector. Provided that*

$r \geq |T \cup \mathrm{supp}(\boldsymbol{x})|$,

$$\|\Phi_T^* \Phi \cdot \boldsymbol{x}|_{T^c}\|_2 \leq \delta_r \|\boldsymbol{x}|_{T^c}\|_2.$$

The last consequence of the restricted isometry property that we will need measures how much the sampling matrix inflates nonsparse vectors. Its proof uses arguments from functional analysis.

Proposition 3 (Energy Bound). *Suppose* $\boldsymbol{\Phi}$ *verifies the upper inequality of* (3), *viz.*

$$\|\Phi \boldsymbol{x}\|_2 \leq \sqrt{1+\delta_r}\, \|\boldsymbol{x}\|_2 \quad \text{when} \quad \|\boldsymbol{x}\|_0 \leq r.$$

Then, for every signal $\boldsymbol{x}$,

$$\|\Phi \boldsymbol{x}\|_2 \leq \sqrt{1+\delta_r}\left[\|\boldsymbol{x}\|_2 + \frac{1}{\sqrt{r}}\|\boldsymbol{x}\|_1\right].$$

5.2. Iteration invariant: sparse case

In proving Theorem 2, we first operate under the assumption that the signal $\boldsymbol{x}$ is exactly s-sparse. We remove this assumption later.

Theorem 3 (Sparse Error Reduction). *Assume* $\boldsymbol{x}$ *is* s*-sparse. For each* $k \geq 0$, *the signal approximation* $\boldsymbol{a}^k$ *is* s*-sparse, and*

$$\|\boldsymbol{x} - \boldsymbol{a}^{k+1}\|_2 \leq 0.5\|\boldsymbol{x} - \boldsymbol{a}^k\|_2 + 7.5\|\boldsymbol{e}\|_2.$$

In particular,

$$\|\boldsymbol{x} - \boldsymbol{a}^k\|_2 \leq 2^{-k}\|\boldsymbol{x}\|_2 + 15\|\boldsymbol{e}\|_2.$$

Remark 3. *This bound assumes the least-squares step in the algorithm is performed without error. In Section 5 of Needell and Tropp,*[19] *we study how many iterations of a least-squares solver are needed to make the resulting error negligible. We show that, if we use conjugate gradient for the estimation step of* CoSaMP, *then initializing the least-squares algorithm with the current approximation* $\boldsymbol{a}^{k-1}$, *then Theorem 3 still holds.*

The proof of the iteration invariant, Theorem 3 involves a sequence of short lemmas, one for each step in the algorithm. We fix an iteration k, and let $\boldsymbol{a} = \boldsymbol{a}^{k-1}$ be the signal approximation at the beginning of the iteration. We define the residual as $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{a}$, which must be $2s$-sparse since both $\boldsymbol{a}$ and $\boldsymbol{x}$ are s-sparse. We also observe that the vector $\boldsymbol{v}$ of updated samples can be interpreted as noisy samples of the residual:

$$\boldsymbol{v} \stackrel{\text{def}}{=} \boldsymbol{u} - \Phi \boldsymbol{a} = \Phi(\boldsymbol{x} - \boldsymbol{a}) + \boldsymbol{e} = \Phi \boldsymbol{r} + \boldsymbol{e}.$$

The first step in the algorithm is the identification step, in which a set of components is found corresponding to locations where the residual signal has a lot of energy. This is summarized by the following lemma which is proven in Needell and Tropp.[19]

Lemma 1 (Identification). *The set* $\Omega = \mathrm{supp}(\boldsymbol{y}_{2s})$, *where* $\boldsymbol{y} = \boldsymbol{\Phi}^* \boldsymbol{v}$ *is the signal proxy, contains at most* $2s$ *indices, and*

$$\|\boldsymbol{r}|_{\Omega^c}\|_2 \leq 0.2223\|\boldsymbol{r}\|_2 + 2.34\|\boldsymbol{e}\|_2.$$

Next, the algorithm merges the support of the current estimate $\boldsymbol{a}$ with the new selection of components. The following simple lemma shows that components of the signal $\boldsymbol{x}$ outside this set have small energy.

Lemma 2 (Support Merger). *Let* Ω *be a set of at most* $2s$ *indices. The set* $T = \Omega \cup \mathrm{supp}(\boldsymbol{a})$ *contains at most* $3s$ *indices, and*

$$\|\boldsymbol{x}|_{T^c}\|_2 \leq \|\boldsymbol{r}|_{\Omega^c}\|_2.$$

The estimation step in the algorithm obtains values for coefficients in the set T by solving a least-squares problem. The next result bounds the error of this approximation.

Lemma 3 (Estimation[19]). *Let* T *be a set of at most* $3s$ *indices, and define the least-squares signal estimate* $\boldsymbol{b}$ *by the formulae*

$$\boldsymbol{b}|_T = \Phi_T^\dagger \boldsymbol{u} \quad \text{and} \quad \boldsymbol{b}|_{T^c} = \boldsymbol{0},$$

where $\boldsymbol{u} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{e}$. *Then*

$$\|\boldsymbol{x} - \boldsymbol{b}\|_2 \leq 1.112\|\boldsymbol{x}|_{T^c}\|_2 + 1.06\|\boldsymbol{e}\|_2.$$

Proof. Note first that

$$\|\boldsymbol{x} - \boldsymbol{b}\|_2 \leq \|\boldsymbol{x}|_{T^c}\|_2 + \|\boldsymbol{x}|_T - \boldsymbol{b}|_T\|_2.$$

Using the expression $\boldsymbol{u} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{e}$ and the fact $\Phi_T^\dagger \Phi_T = \mathbf{I}_T$, we calculate that

$$\begin{aligned}\|\boldsymbol{x}|_T - \boldsymbol{b}|_T\|_2 &= \|\boldsymbol{x}|_T - \Phi_T^\dagger(\Phi \cdot \boldsymbol{x}|_T + \Phi \cdot \boldsymbol{x}|_{T^c} + \boldsymbol{e})\|_2 \\ &\leq \|(\Phi_T^* \Phi_T)^{-1}\Phi_T^* \Phi \cdot \boldsymbol{x}|_{T^c}\|_2 + \|\Phi_T^\dagger \boldsymbol{e}\|_2.\end{aligned}$$

The cardinality of T is at most $3s$, and $\boldsymbol{x}$ is s-sparse, so Proposition 1 and Corollary 1 imply that

$$\|\boldsymbol{x}|_T - \boldsymbol{b}|_T\|_2 \leq \frac{\delta_{4s}}{1-\delta_{3s}}\|\boldsymbol{x}|_{T^c}\|_2 + \frac{\|\boldsymbol{e}\|_2}{\sqrt{1-\delta_{3s}}}.$$

Combine the bounds to reach

$$\|\boldsymbol{x} - \boldsymbol{b}\|_2 \leq \left[1 + \frac{\delta_{4s}}{1-\delta_{3s}}\right]\|\boldsymbol{x}|_{T^c}\|_2 + \frac{\|\boldsymbol{e}\|_2}{\sqrt{1-\delta_{3s}}}.$$

Finally, invoke the hypothesis that $\delta_{3s} \leq \delta_{4s} \leq 0.1$. □

Lastly, the algorithm prunes the intermediate estimation to its largest s terms. The triangle inequality bounds the error in this pruned approximation.

Lemma 4 (Pruning[19]). *The pruned approximation* $\boldsymbol{b}_s$ *satisfies*

$$\|\boldsymbol{x} - \boldsymbol{b}_s\|_2 \leq 2\|\boldsymbol{x} - \boldsymbol{b}\|_2.$$

At the end of the iteration, the new approximation is formed by setting $\boldsymbol{a}^k = \boldsymbol{b}_s$. The sequence of lemmas above allows us to

prove the iteration invariant for the sparse case, Theorem 3. Indeed, we have:

$$\begin{aligned}\|\boldsymbol{x}-\boldsymbol{a}^k\|_2 &= \|\boldsymbol{x}-\boldsymbol{b}_s\|_2\\ &\le 2\|\boldsymbol{x}-\boldsymbol{b}\|_2\\ &\le 2\cdot(1.112\|\boldsymbol{x}|_{T^c}\|_2 + 1.06\|\boldsymbol{e}\|_2)\\ &\le 2.224\|\boldsymbol{r}|_{\Omega^c}\|_2 + 2.12\|\boldsymbol{e}\|_2\\ &\le 2.224\cdot(0.2223\|\boldsymbol{r}\|_2 + 2.34\|\boldsymbol{e}\|_2) + 2.12\|\boldsymbol{e}\|_2\\ &< 0.5\|\boldsymbol{r}\|_2 + 7.5\|\boldsymbol{e}\|_2\\ &= 0.5\|\boldsymbol{x}-\boldsymbol{a}^{k-1}\|_2 + 7.5\|\boldsymbol{e}\|_2.\end{aligned}$$

To obtain the second bound in Theorem 3, we solve the error recursion and observe that

$$(1 + 0.5 + 0.25 + \cdots)\cdot 7.5\|\boldsymbol{e}\|_2 \le 15\|\boldsymbol{e}\|_2$$

This point completes the argument.

5.3. Extension to general signals

We now turn to the proof of the main result for CoSaMP, Theorem A. We must remove the assumption that the signal $\boldsymbol{x}$ is sparse. Although this job seems difficult at first, it turns out to have an elegant solution. We can view the noisy samples of a generic signal as samples of a sparse signal corrupted by additional noise that reflects the tail of the signal. We have the following reduction to the sparse case, of which Theorem 2 is a consequence.

Lemma 5 (Reduction to Sparse Case[19]). *Let $\boldsymbol{x}$ be an arbitrary vector in $\mathbb{C}^N$. The sample vector $\boldsymbol{u} = \boldsymbol{\Phi x} + \boldsymbol{e}$ can also be expressed as $\boldsymbol{u} = \boldsymbol{\Phi x}_s + \tilde{\boldsymbol{e}}$ where*

$$\|\tilde{\boldsymbol{e}}\|_2 \le 1.05\left[\|\boldsymbol{x}-\boldsymbol{x}_s\|_2 + \frac{1}{\sqrt{s}}\|\boldsymbol{x}-\boldsymbol{x}_s\|_1\right] + \|\boldsymbol{e}\|_2.$$

This reduction will now allow us to prove the iteration invariant for general signals, Theorem 2. Let $\boldsymbol{x}$ be a general signal. By Lemma 5, we can write the noisy vector of samples as $\boldsymbol{u} = \boldsymbol{\Phi x}_s + \tilde{\boldsymbol{e}}$. Applying the sparse iteration invariant, Theorem 3, we obtain

$$\|\boldsymbol{x}_s - \boldsymbol{a}^{k+1}\|_2 \le 0.5\|\boldsymbol{x}_s - \boldsymbol{a}^k\|_2 + 7.5\|\tilde{\boldsymbol{e}}\|_2.$$

We then invoke the lower and upper triangle inequalities to obtain

$$\|\boldsymbol{x} - \boldsymbol{a}^{k+1}\|_2 \le 0.5\|\boldsymbol{x}-\boldsymbol{a}^k\|_2 + 7.5\|\tilde{\boldsymbol{e}}\|_2 + 1.5\|\boldsymbol{x}-\boldsymbol{x}_s\|_2.$$

Finally, recalling the estimate for $\|\tilde{\boldsymbol{e}}\|_2$ from Lemma 5 and simplifying, we reach

$$\begin{aligned}\|\boldsymbol{x}-\boldsymbol{a}^{k+1}\|_2 &\le 0.5\|\boldsymbol{x}-\boldsymbol{a}^k\|_2 + 9.375\|\boldsymbol{x}-\boldsymbol{x}_s\|_2\\ &\quad + \frac{7.875}{\sqrt{s}}\|\boldsymbol{x}-\boldsymbol{x}_s\|_1 + 7.5\|\boldsymbol{e}\|_2\\ &< 0.5\|\boldsymbol{x}-\boldsymbol{a}^k\|_2 + 10\nu.\end{aligned}$$

where ν is the unrecoverable energy (Equation 4). C

References

1. Candès E, Romberg J., Tao T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete Fourier information. *IEEE Trans. Info. Theory 52*, 2 (Feb. 2006), 489–509.
2. Candès, E.J. The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris, Serie I 346* (2008), U589–U592.
3. Candès, E.J., Romberg, J., Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pur. Appl. Math. 59*, 8 (2006), 1207–1223.
4. Candès, E.J., Tao, T. Decoding by linear programming. *IEEE Trans. Info. Theory 51* (2005), 4203–4215.
5. Candès, E.J., Tao, T. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Info. Theory 52*, 12 (Dec. 2006), 5406–5425.
6. Cohen, A., Dahmen, W., DeVore, R. Compressed sensing and best k-term approximation. *J. Am. Math. Soc. 22*, 1 (2009) 211–231.
7. Cormen, T., Lesierson, C., Rivest, L., Stein, C. *Introduction to Algorithms*, 2nd edition, MIT Press, Cambridge, MA, 2001.
8. Cormode, G., Muthukrishnan, S. Combinatorial algorithms for compressed sensing. Technical report, DIMACS, 2005.
9. Dai, W., Milenkovic, O. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Info. Theory 55*, 5 (2009).
10. Donoho, D.L. Compressed sensing. *IEEE Trans. Info. Theory 52*, 4 (Apr. 2006), 1289–1306.
11. Donoho, D.L., Stark, P.B. Uncertainty principles and signal recovery. *SIAM J. Appl. Math. 49*, 3 (June 1989)906–931.
12. Donoho, D.L., Tsaig, Y., Drori, I., Starck, J.-L. Sparse solution of underdetermined linear equations by stagewise Orthogonal Matching Pursuit (StOMP). Submitted for publication (2007)
13. Foucart, S. A note on guaranteed sparse recovery via ℓ_1-minimization. *Appl. Comput. Harmon. Anal.* (2010). To appear.
14. Gilbert, A., Strauss M., Tropp J., Vershynin R. Algorithmic linear dimension reduction in the ℓ_1 norm for sparse vectors. In *Proceedings of the 44th Annual Allerton Conference on Communication, Control and Computing* (Sept. 2006).
15. Gilbert, A., Strauss, M., Tropp, J., Vershynin, R. One sketch for all: fast algorithms for compressed sensing. In *Proceedings of the 39th ACM Symposium. Theory of Computing* (San Diego, June 2007).
16. Gilbert, A.C., Guha, S., Indyk, P., Muthukrishnan, S., Strauss, M.J. Near-optimal sparse Fourier representations via sampling. In *Proceedings of the 2002 ACM Symposium on Theory of Computing STOC* (2002), 152–161.
17. Kim, S.-J., Koh, K., Lustig, M., Boyd, S., Gorinevsky, D. An interior-point method for large-scale ℓ_1-regularized least squares. *IEEE J. Sel. Top. Signa. 1*, 4 (2007) 606–617.
18. Needell, D., Tropp, J.A. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. ACM Technical Report 2008–01, California Institute of Technology, Pasadena, July 2008.
19. Needell, D., Tropp, J.A. CoSaMP: Iterative signal recovery from noisy samples. *Appl. Comput. Harmon. Anal. 26*, 3 (2008), 301–321.
20. Needell, D., Vershynin, R. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *IEEE J. Sel. Top. Signa.* (2007). To appear.
21. Nesterov, Y.E., Nemirovski, A.S. *Interior Point Polynomial Algorithms in Convex Programming*. SIAM, Philadelphia, 1994.
22. Paige, C.C., Saunders, M.A. LSQR: An algorithm for sparse linear equations and sparse least squares. *TOMS 8*, 1 (1982), 43–71.
23. Rudelson, M., Vershynin, R. Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements. In *Proceedings of the 40th Annual Conference on Info. Sciences and Systems*, Princeton, Mar. 2006.
24. Tropp, J.A., Gilbert, A.C. Signal recovery from random measurements via Orthogonal Matching Pursuit. *IEEE Trans. Info. Theory 53*, 12 (2007), 4655–4666.

Deanna Needell (dneedell@stanford.edu), Stanford University, Stanford, CA.

Joel A. Tropp (jtropp@acm.caltech.edu), California Institute of Technology, Pasadena, CA.

DOI:10.1145/1859204.1859230

Technical Perspective
QIP = PSPACE Breakthrough

By Scott Aaronson

QUANTUM COMPUTERS ROCKETED to public attention—or at least to the attention of a specific part of the public sector—in the mid-1990s with the discovery that a computer operating on quantum principles could solve certain problems exponentially faster than we know how to solve them with computers today. The most famous of these problems is factoring large numbers, a feat that would enable one to break most of the cryptography currently used on the Internet. While quantum computers large enough to do anything useful haven't been built yet, the theory of quantum computing has developed rapidly.

Researchers have discovered quantum algorithms for a variety of problems, such as searching databases and playing games. However, it is now clear that for a wide range of problems, quantum computers offer little or no advantage over their classical counterparts.

The following paper describes a breakthrough result that ives a very general situation in which quantum computers are no more useful than classical ones. The result settles a longstanding problem about *quantum interactive proof systems* showing they are no more (or less) powerful than classical interactive proof systems.

What *is* an interactive proof system? Basically, it's an imagined process in which a prover (named Merlin) tries to convince a skeptical verifier (named Arthur) that a mathematical statement is true, by submitting himself to interrogation. Merlin, though untrustworthy, has unlimited computational powers; Arthur, by contrast, is limited to performing computations that take polynomial time. By asking Merlin pointed questions, Arthur can sometimes convince himself of a statement more quickly than by reading a conventional proof.

When confronted with a new model of computation, theoretical computer scientists' first instinct is to name the model with an inscrutable sequence of capital letters. And thus, in 1985, Goldwasser, Micali, and Rackoff as well as Babai defined the complexity class IP (Interactive Proofs), which consists of all mathematical problems for which Merlin can convince Arthur of a "yes" answer by a probabilistic, interactive protocol. They then asked: *how big is* IP? In a dramatic development in 1990, Lund et al. and Shamir showed that IP was larger than almost anyone had imagined.

Using an innovative argument based on polynomials over finite fields, the authors showed that IP contains PSPACE (Polynomial Space), the class of problems solvable by a classical computer using a polynomial amount of memory but possibly an exponential amount of time. PSPACE is known to encompass games of strategy, such as chess and Go. And thus, we get the surprising implication that, if aliens with infinite computational powers came to earth, they could not only beat humans at chess, but could also *mathematically prove they were playing chess perfectly*. Since it's not difficult to show that every IP problem is also a PSPACE problem, we obtain one of the most famous equations in CS: IP = PSPACE. This equation paved the way for many advances of a more down-to-earth nature: for example, in cryptography and program checking.

When quantum computing first came along, almost every topic in theoretical CS was ripe for reexamination in light of quantum effects. In 1999, Kitaev and Watrous defined the complexity class QIP (Quantum Interactive Proofs), which is like IP except that now Arthur and Merlin both have quantum computers at their disposal, and can send and receive quantum messages.

Because of the 'exponential parallelism' inherent in quantum mechanics—a state of n quantum bits (or qubits) requires a vector of 2^n complex numbers to describe—it was a reasonable guess that sending and receiving quantum messages would let Arthur verify even more mathematical statements than he could using classical interaction. In the beginning, though, all that seemed clear was that quantum interactive proofs were at least as powerful as classical ones: IP = PSPACE $\subseteq$ QIP.

The reason is what I call the "Clark Kent principle": like Superman concealing his powers, a quantum computer can always "hide its quantumness" and emulate a classical computer if necessary. So the real question is whether QIP is *larger* than IP. Kitaev and Watrous managed to show that QIP was contained in EXP, the class of problems solvable by a classical computer using an exponential amount of time. Since EXP is believed to be strictly larger than PSPACE, this put QIP somewhere in a no-man's-land between PSPACE and EXP.

The authors have finally pinned down the power of quantum interactive proofs: they show that QIP = IP = PSPACE. In other words, quantum interactive proofs have exactly the same power as classical interactive proofs: both of them work for all problems in PSPACE but no other problems. In proving this, the authors confronted an extremely different challenge than that confronted in the earlier IP = PSPACE proof. Instead of demonstrating the *power* of interactive proofs, the authors had to show that quantum interactive proofs were *weak* enough to be simulated using polynomial memory: that is, QIP $\subseteq$ PSPACE.

To achieve this, the authors use a powerful recent tool called the *multiplicative weights update method*. Interestingly, computer scientists originally developed this method for reasons having nothing to do with quantum computing and, completing the circle, the QIP = PSPACE breakthrough is already leading to new work on the classical applications of the multiplicative weights method. This illustrates how advances in quantum and classical computing are sometimes increasingly difficult to tell apart. C

Scott Aaronson is an associate professor of Electrical Engineering and Computer Science at MIT, Cambridge, MA.

DOI:10.1145/1859204.1859231

QIP = PSPACE

By Rahul Jain, Zhengfeng Ji, Sarvagya Upadhyay, and John Watrous

Abstract

The interactive proof system model of computation has been studied extensively in computational complexity theory and theoretical cryptography for more than 25 years, and has driven the development of interesting new techniques and insights in those fields. This work considers the quantum interactive proof system model, which is the classical model's natural quantum computational analog. An exact characterization of the expressive power of quantum interactive proof systems is obtained: the collection of computational problems having quantum interactive proof systems consists precisely of those problems solvable with an ordinary classical computer using at most a polynomial amount of memory (or QIP = PSPACE in complexity-theoretic terminology). One striking implication of this characterization is that it implies quantum computing provides no increase in computational power whatsoever over classical computing in the context of interactive proof systems.

1. INTRODUCTION

The notion of a *proof* has fundamental importance in the theory of computation. Indeed, the foundational work of Alonzo Church and Alan Turing in the 1930s, which produced the first formal models of computation (λ-calculus and Turing machines), was principally motivated by questions concerning proofs in formal logic. The theory of NP-completeness, developed in the 1970s by Stephen Cook, Leonid Levin, and Richard Karp, provides another example. It is built on the notion of *efficient* proof verification, and is arguably the most widely applicable discovery ever made in the theory of computation.

This paper is concerned with the potential advantages offered by *quantum computation* in the setting of proofs, and in particular its advantages when applied to the *interactive proof system* model of computation. Considered by many to be a cornerstone of modern computational complexity theory, the interactive proof system model was first introduced in the mid-1980s, and its quantum computational variant has been an object of study in quantum computing for more than a decade.

The main result to be presented herein is that quantum computation does not enhance the expressive power of interactive proof systems at all: quantum and classical interactive proof systems are equivalent in power, both coinciding with the complexity class PSPACE of computational problems solvable using an amount of memory scaling polynomially in the length of the input to the problem. This resolves a fundamental question about the quantum interactive proof system model that has been open since its introduction.

1.1. Randomness and interaction in proofs

When speaking of proofs, one typically has a traditional notion in mind where, at least in an abstract sense, the proof itself is a string of symbols to be verified by some form of computation. In the theory of NP-completeness, a proof is generally a string of bits certifying that a given object has a property of interest, such as a graph having a three-coloring. In mathematics, proofs appear in journal papers and books, to be verified by mathematicians with interest in the claimed theorems. Although it is not done, one imagines that in principle—and undoubtedly through monumental effort—such proofs could be transformed into a purely symbolic form verifiable by a computer, presumably in accordance with axiomatic set theory.

There are, however, other interesting notions of proofs that extend the traditional notion in various ways. For instance, *randomness* may be used in the verification process to gather statistical evidence for or against a given claim. For example, if a coin is flipped 1,000 times, and heads come up 975 times, one can be reasonably sure that the coin is biased toward heads—and although the claim has not been proved in the sense of formal logic, it would be foolish to hold out any hope that the coin is actually fair. Allowing for an *interaction* in the process of verification, where the proof takes the form not of a static string of symbols, but as a process that may receive input and produce output, is another extension. The *interactive proof system* model, first proposed by Shafi Goldwasser et al.[8] and László Babai[2] in 1985, combines both the extensions of randomness and interaction into a formal computational model.

One variant of a well-known, often-repeated, and highly informal example illustrating a benefit of interaction and randomness in proofs is as follows. Suppose that you can taste the difference between Coke and Pepsi, and that you wish to convince a person that is skeptical of this claim. You cannot reasonably hope, for a variety of reasons, to succeed in this task through the traditional notion of a proof, but it is easily achieved through the use of interaction and randomness. In particular, you may allow the skeptic to subject you to a "blind taste test," where you are offered Coke or Pepsi in an unmarked can—and when you are able to repeatedly make the correct identification, over the course of many independent random tests, the skeptic should be convinced that you can taste the difference. The negligible—but nevertheless nonzero—probability that every single identification was made through luck rather than an actual difference in taste is accepted: the formal definition of the model requires a high probability, but not absolute certainty, of correct outcomes.

As a computational model, interactive proof systems are not typically considered for single, isolated statements such

The original version of this paper was published in the *42nd ACM Symposium on Theory of Computing*, 2010.

as "Coke and Pepsi taste different"—irrespective of that example's informality. Rather, as is done in the theory of NP-completeness, interactive proof systems are connected to computational *decision problems* where input strings are to be classified as *yes*-inputs and *no*-inputs. A particular decision problem is said to have an interactive proof system if there exists a *computationally efficient* verification procedure (called the *verifier*) with two properties that capture the essence of what it means to be a proof:

1. *Completeness*. This property represents the requirement that true statements can be proved. Here, the requirement is that for any *yes*-input string x, there exists a behavior of the entity the verifier interacts with (called the *prover*) that causes the verifier to believe that x is indeed a *yes*-input. This situation may be indicated by the verifier outputting 1 (or *accept*) after interacting with the prover.
2. *Soundness*. This property represents the complementary requirement to completeness, which is that false statements cannot be proved. In the current situation, the requirement is that the probability that the verifier will be convinced to output 1 given a *no*-input is negligible, regardless of the prover's actions. Instead, the verifier outputs 0 (or *reject*) with probability very close to 1, indicating the prover's failure to convince it otherwise.

While the verifier is restricted to be computationally efficient (or more formally to be describable as a polynomial-time probabilistic Turing machine), no such restriction is placed on the prover. These assumptions serve to place an emphasis on *efficient verification*, as opposed to *efficient construction* of proofs.

It must be stressed that there is an inherent asymmetry between the completeness and soundness conditions: when the verifier outputs 0 (or *reject*), it is not necessarily convinced that the input is a *no*-input, but only that the prover has failed to convince it that the input is a *yes*-input. This is analogous to the traditional notion of a proof: one would not be convinced that a particular mathematical statement is false by seeing an incorrect proof claiming it is true.

The most fundamental question, from the viewpoint of complexity theory, about the interactive proof system model is: which computational decision problems have interactive proof systems? The answer is known: a decision problem has an interactive proof system if and only if it is solvable by an ordinary computer (or deterministic Turing machine) that requires an amount of memory that scales at most polynomially in its input length. A more succinct expression of this fact is given by the equation IP = PSPACE. In this equation, IP denotes the set of decision problems having interactive proof systems, PSPACE represents the set of decision problems solvable using polynomial memory (or *space*), and of course the equality expresses the fact that the two sets are equal, meaning that the two descriptions give rise to precisely the same set of problems.

Like all set equalities, there are two subset relations represented by the equation IP = PSPACE. One of the two relations, IP $\subseteq$ PSPACE, is easy to prove: the typical proof involves a fairly straightforward recursive traversal of a *game tree* whose edges represent messages exchanged between the prover and verifier, which can be performed in a space-efficient way. The other relation, PSPACE $\subseteq$ IP, was proved by Adi Shamir[14] in 1990, based on work of Carsten Lund et al.[10] It is a landmark result that established the powerful proof technique of *arithmetization* as a standard tool in computational complexity.

1.2. Quantum computation in proofs

The idea of *quantum computation* was born in the early 1980s when Richard Feynman[7] asked a brilliant question: If quantum mechanics is so hard to simulate with a classical computer, why not build a computer based on quantum mechanics to simulate quantum mechanics more directly? Feynman's ideas on the subject led David Deutsch[6] to define the *quantum Turing machine* model and to investigate its computational power. Driven in large part by Peter Shor's subsequent discoveries of polynomial-time algorithms for factoring and computing discrete logarithms on a quantum computer,[15] quantum computation has developed into an active field of study within theoretical computer science and both theoretical and experimental physics.

Large-scale quantum computers do not currently exist, and it is universally agreed that at the very least their realization will be an enormous technological challenge. However, it must also be appreciated that quantum mechanics is a remarkably accurate theory that has never been refuted—and with the theory suggesting that quantum computing should be possible, scientists are naturally compelled to test the theory by attempting to build a quantum computer. Efforts to do this are underway in many laboratories around the world.

Within the theoretical study of quantum computation, it is natural to consider quantum computational variants of interesting classical models, including those based on the notion of proofs. The *quantum interactive proof system model*, which was first introduced in 1999,[17] represents a natural quantum computational analog to the (classical) interactive proof system model. In simple terms, the quantum model allows the verifier and prover in an interactive proof system to perform quantum computations and exchange quantum information, but is otherwise similar to the classical model.

The potential advantages of quantum computation in the setting of interactive proof systems are not limited to the fact that the verifier is able to perform a potentially wider range of computations. The nature of quantum information is such that it has striking benefits in a variety of information-processing tasks, such as secret key exchange[3] and distributed computational tasks allowing limited communication.[13] A known benefit of quantum computation in the interactive proof system model is that it allows for a major reduction in the *number of messages* that must be exchanged: quantum interactive proof systems allowing just three messages to be exchanged between the prover and verifier have the full power of those allowing any polynomial number of messages.[9]

It is not known if the analogous fact holds for classical interactive proof systems, but it is conjectured not to hold. (It would, in particular, send complexity theorists reeling from the collapse of the polynomial-time hierarchy if it were true.)

It is not difficult to show that quantum interactive proof systems are at least as powerful as classical ones—due, in essence, to the fact that quantum computers can mimic classical computers. This implies PSPACE $\subseteq$ QIP. Unlike the subset relation IP $\subseteq$ PSPACE, however, it is not straightforward to prove QIP $\subseteq$ PSPACE, and prior to the work presented in this paper this relationship was not known. The remainder of this paper is devoted to a presentation of this result.

2. QUANTUM PROOF SYSTEMS

This section aims to provide readers with a basic understanding of quantum information and the quantum interactive proof system model, narrowly focused on material that is required for the subsequent sections of this paper. The standard text, Nielsen and Chuang,[12] is recommended to those readers interested in a more comprehensive presentation of the much more broad field of quantum information and quantum computation.

2.1. States and measurements

Quantum information is described in the language of matrices and linear algebra in way that is reminiscent of probabilistic models such as Markov chains.

Consider a physical device X whose possible states are the binary strings of length k for some fixed choice of a positive integer k. For brevity, one may say that X is a *k-bit register*, or a *k-qubit register* in the quantum setting, as a way of suggesting that X is a device used for storing and processing information.

One way to represent one's knowledge of X in a probabilistic setting, where the states are changed through a randomized process of some sort, is by a vector v of probabilities: the value $v[x]$ represents the probability that X takes the state x for each binary string x of length k. The vector v is therefore a K-dimensional vector for $K = 2^k$.

In quantum information, the K-dimensional vector v of probabilities is replaced by a $K \times K$ matrix ρ with complex number entries, known as a *density matrix*. (It is traditional in quantum physics to use lower-case Greek letters, often ρ, σ, and ξ, to represent density matrices.) It is reasonable to view that the diagonal entries of a density matrix ρ represent probabilities, so that a "standard measurement" of the register X would yield each possible state x with probability $\rho[x, x]$. The off-diagonal entries of ρ are not easily connected to classical intuitions, but they do have great significance with respect to their role in calculations. Informally speaking, for distinct choices of binary strings x and y, the off-diagonal entries $\rho[x, y]$ and $\rho[y, x]$ provide information about the degree to which the states x and y are in "superposition" in X, or alternately the degree to which these states could *interfere* with one another in processes involving X.

Although they are not as intuitive as vectors of probabilities, density matrices are very simple objects in a mathematical sense: they are *positive semidefinite* matrices whose diagonal entries sum to 1 (i.e., whose *trace* is 1). A matrix ρ is positive semidefinite if and only if it satisfies (i) the condition that $\rho[x, y] = \overline{\rho[y, x]}$ for all choices of x and y (with $\overline{\alpha}$ denoting the complex conjugate of α), and (ii) the constraint that all of its eigenvalues are nonnegative.

Quantum states of independent registers are represented by *tensor products* of density matrices. For instance, if X and Y are k-qubit registers independently prepared in the quantum states represented by the $K \times K$ density matrices σ and ξ, then the quantum state of the pair (X, Y) is described by the $K^2 \times K^2$ density matrix

$$\sigma \otimes \xi = \begin{pmatrix} \sigma[0,0]\xi & \cdots & \sigma[0,K-1]\xi \\ \vdots & \ddots & \vdots \\ \sigma[K-1,0]\xi & \cdots & \sigma[K-1,K-1]\xi \end{pmatrix}$$

where the binary strings of length k indexing the entries of σ have been indicated by the integers they represent in binary notation.

Of course, not every state of a pair of registers (X, Y) can be expressed as a tensor product in this way, representing the fact that there may be dependencies between X and Y. Two such examples for the case $k = 1$ are the following density matrices:

$$\rho_0 = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \end{pmatrix} \quad \text{and} \quad \rho_1 = \begin{pmatrix} \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{pmatrix}.$$

The first density matrix ρ_0 represents a simple situation in which X and Y are randomly correlated: the two registers take the same state, chosen to be 0 or 1 uniformly at random. The second density matrix ρ_1 is similar to ρ_0 in the sense that it represents a joint state with a strong correlation between X and Y. It so happens, however, that the two nonzero off-diagonal entries endow ρ_1 with the characteristic of *entanglement*, which is one of the most remarkable features of quantum information. (*Quantum teleportation*[4] provides a well-known example of the use of the entangled state ρ_1 as a resource in an information-processing task.)

For every possible quantum state of a pair of registers (X, Y), whether correlated or not, there is a uniquely determined *reduced state* of the register X that, in essence, would describe the state of X if Y were to suddenly be destroyed or lost. This is analogous to the *marginal* probability distribution of X in the probabilistic case. In mathematical terms the reduced state is defined by an operation known as the *partial trace*. If it is the case that the state of the pair (X, Y) is described by a $K^2 \times K^2$ density matrix ρ, which may be written as a *block matrix* of the form

$$\rho = \begin{pmatrix} M_{0,0} & \cdots & M_{0,K-1} \\ \vdots & \ddots & \vdots \\ M_{K-1,0} & \cdots & M_{K-1,K-1} \end{pmatrix}$$

for $K \times K$ matrices $M_{i,j}$, then the $K \times K$ reduced density matrix for the register X is defined as

$$\text{PartialTrace}\begin{pmatrix} M_{0,0} & \cdots & M_{0,K-1} \\ \vdots & \ddots & \vdots \\ M_{K-1,0} & \cdots & M_{K-1,K-1} \end{pmatrix} = \begin{pmatrix} \text{Trace}(M_{0,0}) & \cdots & \text{Trace}(M_{0,K-1}) \\ \vdots & \ddots & \vdots \\ \text{Trace}(M_{K-1,0}) & \cdots & \text{Trace}(M_{K-1,K-1}) \end{pmatrix}.$$

For the states ρ_0 and ρ_1 defined previously, for example, it holds that

$$\text{PartialTrace}(\rho_0) = \text{Partial Trace}(\rho_1) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

One can consider other variants of this notion, such as one defining the reduced state of Y rather than X or one for the situation when X and Y have different sizes, but the case presented above is sufficient for the needs of this paper.

Now suppose one has a k-qubit register X in their possession, and that the state of X is described by the density matrix ρ. Much like the probabilistic case, there is a limit to the amount of information about ρ that can be gained by looking at a single copy of X. Any information must be obtained by a *measurement*, which will generally result in a probabilistic outcome.

In more precise terms, a measurement of X that results in one of the outcomes 0, 1,..., $m-1$ (for some choice of m) is described by a collection of positive semidefinite matrices $\{\Pi_0, \Pi_1, ..., \Pi_{m-1}\}$ that satisfy the condition $\Pi_0 + \cdots + \Pi_{m-1} = \mathbb{1}$, where $\mathbb{1}$ represents the $K \times K$ identity matrix. Often the matrices $\Pi_0, ..., \Pi_{m-1}$ are *projection matrices*, meaning that in addition to being positive semidefinite their only eigenvalues are 0 and 1. This is not a requirement, but every measurement considered in this paper may be assumed to have this property.

Now, when ρ is measured with respect to the measurement $\{\Pi_0, ..., \Pi_{m-1}\}$, each outcome $a \in \{0, ..., m-1\}$ is obtained with probability $\text{Trace}(\Pi_a \rho)$. The condition that ρ and Π_a are positive semidefinite implies that the resulting probabilities are nonnegative, and the conditions that $\text{Trace}(\rho) = 1$ and $\Pi_0 + \cdots + \Pi_{m-1} = \mathbb{1}$ imply that the probabilities sum to 1. According to the simplest definition, the register X is destroyed by the act of measurement, so there is no opportunity to gain further information about ρ unless additional independent copies of it are made available.

One example of a measurement, referred to as the "standard measurement" in passing above, is obtained by taking $m = K$, associating the possible outcomes 0,..., $K-1$ with the length k binary strings with respect to binary notation, and taking Π_x to be the matrix with a single 1 in the (x, x) diagonal entry and 0 for all other entries. As suggested before, this measurement results in each outcome x with probability $\rho[x, x]$. The measurements that will be of interest in this paper have a somewhat different nature, in that they are binary-valued measurements resulting from hypothetical quantum computations performed on registers. They are, nevertheless, still describable within the framework just explained.

One additional component of quantum information theory that has not been mentioned above is the description of *transformations* of quantum registers according to physical processes (such as computations). These changes are described mathematically by linear mappings known as *channels*, and are of great interest to the theory. It is, however, not necessary to make use of this notion in this paper, so it is not discussed further.

2.2. Quantum interactive proofs simplified

Taken in its most general form, the quantum interactive proof system model can allow for complicated and mathematically unwieldy interactions involving the exchange of quantum messages over the course of many rounds. By the nature of quantum information, such an interaction cannot generally be described by the sort of game tree that describes a classical interaction—the possibility of entanglement among the prover, verifier, and message registers prohibits this.

It is, however, always possible[9, 11] to transform a given quantum interactive proof system into one with the following conceptually simple form:

1. The prover sends the verifier a k-qubit register X, for some choice of a positive integer k—which must be polynomially related to the length of the input string x. Upon receiving X, the verifier sets it aside without interacting with it.
2. The verifier chooses a bit $a \in \{0, 1\}$, uniformly at random, and sends a to the prover.
3. The prover sends the verifier a second k-qubit register Y. The state of Y may of course be dependent on the random bit a, given that the prover learned a before sending Y to the verifier, but its correlation with X is limited by the fact that the prover did not have access to X after seeing a.
4. The verifier measures the pair (X, Y) with respect to one of two binary-outcome measurements: $\{\Pi_0^0, \Pi_1^0\}$ in case $a = 0$ and $\{\Pi_0^1, \Pi_1^1\}$ in case $a = 1$. The measurement outcome is interpreted as the verifier's output: 1 means the proof is *accepted*, 0 means it is *rejected*.

The verifier's measurements $\{\Pi_0^0, \Pi_1^0\}$ and $\{\Pi_0^1, \Pi_1^1\}$ will, naturally, be dependent on the input string x to the decision problem under consideration. In accordance with the definition of the quantum interactive proof system model, these measurements must be efficiently implementable by a quantum computation to be performed by the verifier.

It is beyond the scope of this paper to describe how a general quantum interactive proof system can be transformed into one having the above form, but it is not overly complex. The transformation is such that the verifier's measurements can be forced to output 1 with certainty when the input is a *yes*-input, while the maximum probability for the output 1 can be made arbitrarily close to 1/2 when the input is a *no*-input. (More formally speaking, the transformation can

guarantee a probability of at most $1/2 + \varepsilon$ for any fixed choice of a constant $\varepsilon > 0$.)

A common question about quantum interactive proof systems having the above form is this: Why cannot the prover simply prepare three registers X, Y_0, and Y_1, send all three to the verifier in a single message, and allow the verifier to measure (X, Y_0) or (X, Y_1) depending on its choice of the random bit a? This would seem to eliminate the need for interaction, as the verifier would never send anything to the prover. The reason is that entanglement prevents this from working: in order for the two measurements to result in the output 1 with high probability, the registers X and Y will generally need to be highly entangled. One of the curious features of entanglement, however, is that any single quantum register is highly constrained with respect to the degree of entanglement it may simultaneously share with two or more other registers. (This phenomenon was colorfully named the *monogamy of entanglement* by Charles Bennett.) Thus, again in the general case, the only way for the prover to cause the verifier to output 1 is to prepare X in a highly entangled state with a register of its own, and then use this entanglement to prepare Y once the random bit a has been received.

There are many strategies that a prover could potentially employ in an attempt to cause a verifier to output 1 in the type of proof system described above—but they can all be accounted for by considering the possible states of the pair (X, Y) that the verifier measures, conditioned on the two possible values of the random bit a. That is, for any two $K^2 \times K^2$ density matrices ρ_0 and ρ_1, one may ask whether it is possible for the prover to follow a strategy that will leave the verifier with the state ρ_0 for the pair (X, Y) when the random bit takes the value $a = 0$, and with the state ρ_1 when $a = 1$; and the set of all possible such choices for ρ_0 and ρ_1 is simple to characterize. It is precisely the set of all $K^2 \times K^2$ density matrices ρ_0 and ρ_1 for which

$$\text{PartialTrace}(\rho_0) = \text{PartialTrace}(\rho_1). \quad (1)$$

The necessity of this condition follows immediately from the fact that the prover cannot touch X at any point after learning a, while the sufficiency of the condition requires an analysis based on standard mathematical tools of quantum information theory.

It follows that the maximum probability for a prover to cause the verifier to output 1 in the type of proof system described above is the maximum of the quantity

$$\frac{1}{2}\,\text{Trace}\,(\Pi_1^0 \rho_0) + \frac{1}{2}\,\text{Trace}\,(\Pi_1^1 \rho_1) \quad (2)$$

subject to the conditions that ρ_0 and ρ_1 are $K^2 \times K^2$ density matrices satisfying Equation 1.

3. PSPACE AND BOOLEAN CIRCUITS

Based on the claims of the previous section, the characterization QIP = PSPACE follows from the existence of a polynomial-space algorithm for approximating the prover's optimal success probability in a quantum interactive proof system of the highly restricted form described above. The accuracy of this approximation may, in fact, be very coarse: it is only necessary for the algorithm to distinguish an optimal probability of 1 from one very close to 1/2.

The design of space-bounded algorithms is an unnatural task for most algorithm designers. When one cares only about space-efficiency, and not about time-efficiency, programming techniques that would be ridiculous in a practical situation become useful. For example, rather than storing bits used frequently in a given computation, one may instead opt to recompute and then discard those bits every time they are used. Taking such schemes to an extreme, it becomes possible to implicitly perform computations on numbers and matrices that are themselves exponentially larger than the total memory used for the entire computation.

Fortunately, in the late 1970s, Alan Borodin described a simple way to translate the task of highly space-efficient algorithm design into one that is arguably much more natural and intuitive for algorithm designers.[5] For the particular case of PSPACE algorithms, Borodin's result states that a given decision problem is in PSPACE if and only if it can be computed by a collection of Boolean circuits having the form illustrated in Figure 1. The primary appeal of this reformulation is that it allows one to make use of extensive work on *parallel algorithms* for performing various computational tasks when designing PSPACE algorithms.

Now, the task at hand is to prove that any decision problem having a quantum interactive proof system can be decided

Figure 1. A decision problem is solvable in polynomial space if and only if it can be computed by a family of Boolean circuits, one for each input length *n*, whose size may be exponential in *n* but whose depth is polynomial in *n*. (The circuits must also satisfy a *uniformity constraint* that limits the computational difficulty of computing each circuit's description.)

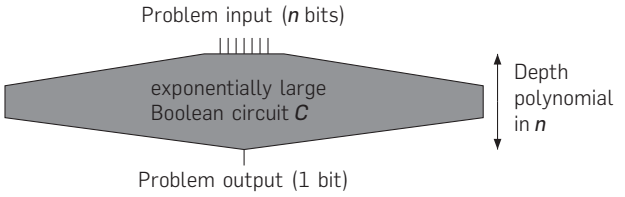
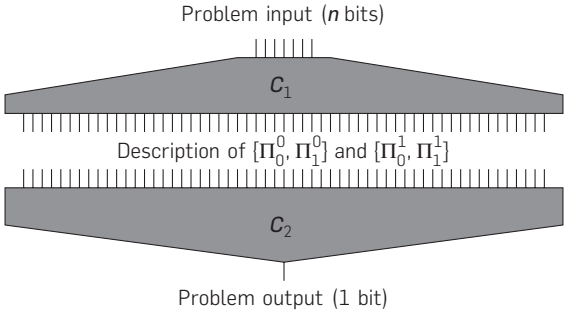


Figure 2. An illustration of the two-stage process for solving any problem in QIP by a bounded-depth Boolean circuit. The circuit C_1 transforms the input string into an exponentially large description of the verifier's measurements $\{\Pi_0^0, \Pi_1^0\}$ and $\{\Pi_0^1, \Pi_1^1\}$, and the circuit C_2 implements a *parallel* algorithm for approximating the optimal value of the associated optimization problem.



by a family of polynomial-depth circuits. There is a natural two-stage process associated with this task that is illustrated in Figure 2. The first stage of this process turns out to be fairly straightforward, using elementary facts about quantum computations and bounded-depth circuits. The second stage is more difficult, and the algorithm for accomplishing it is representative of the main technical contribution of this work. The remainder of the paper is devoted to a discussion of this algorithm.

4. THE MAIN ALGORITHM

The algorithm corresponding to the second stage of the computation illustrated in Figure 2 will now be described. As is required to prove QIP = PSPACE, it is a *highly parallel* algorithm for approximating the value of the optimization problem described at the end of Section 2. This optimization problem is an example of a *semidefinite program* (or SDP for short): it asks for the maximum value of a linear function of a collection of matrix variables that are subject to a collection of linear and positive semidefinite constraints.

SDPs are well studied and subject to an analytically powerful *duality theory* that will be useful in the section following this one. There do exist efficient algorithms to approximate solutions to most SDPs, including the theoretically important *ellipsoid method* and the more practical family of *interior point* methods. Unfortunately these algorithms do not parallelize, and there is a generally-believed complexity-theoretic conjecture (namely NC ≠ P) that implies that no other general and efficient algorithm will parallelize either: some SDPs seem to force algorithms that approximate them to be inherently sequential. For this reason, the algorithm to be presented below will, presumably by necessity, exploit the specific nature of the SDP at hand to allow for its parallelizability. It does not approximate solutions to general SDPs, only those arising from the specific type of quantum interactive proof system described in Section 2.

The algorithm employs a method from learning theory and combinatorial optimization sometimes known as the *matrix multiplicative weights update method*, and in particular makes use of a variant of this technique developed by Manfred Warmuth and Dima Kuzmin[16] and Sanjeev Arora and Satyen Kale.[1]

The algorithm is iterative in nature: for successive values of $t = 0, 1, 2, \ldots$, the algorithm will generate $K^2 \times K^2$ matrices $\rho_0^{(t)}$ and $\rho_1^{(t)}$ that, roughly speaking, correspond to guesses for ρ_0 and ρ_1 in the SDP. In addition, the algorithm also generates a $K \times K$ matrix $\sigma^{(t)}$ for each t, which intuitively represents a common "target" state for

$$\text{PartialTrace}\left(\rho_0^{(t)}\right) \quad \text{and} \quad \text{PartialTrace}\left(\rho_1^{(t)}\right).$$

Like many iterative algorithms, there is a balance between the algorithm's rate of convergence and accuracy. It is essential for the parallelizability of the algorithm that $\rho_0^{(t)}, \rho_1^{(t)}$, and $\sigma^{(t)}$ converge quickly to choices that reveal a near-optimal solution; but for the sake of the algorithm's accuracy it must not move too rapidly in any one direction, for fear of overreacting to poor choices that had the appearance of good ones within a small region of the search space.

To aid in this balance, it is helpful at a technical level to introduce two $K^2 \times K^2$ *penalty matrices*

$$P_0 = \Pi_1^0 + \alpha\Pi_0^0 \quad \text{and} \quad P_1 = \Pi_1^1 + \alpha\Pi_0^1,$$

which have the effect of "inflating" density matrices ρ_0 and ρ_1 that would not yield a value of the objective function (2) close to 1. With a bit of algebra it can be shown that, for a sufficiently large value of $\alpha > 0$, the optimal value of the original SDP will be close to the maximum value of

$$\text{Trace}\left(\frac{1}{2}Q_0 + \frac{1}{2}Q_1\right)$$

over all choices of positive semidefinite matrices Q_0 and Q_1, subject to the constraint that σ – PartialTrace $(P_0 Q_0 P_0)$ and σ – PartialTrace $(P_1 Q_1 P_1)$ are positive semidefinite for some choice of a density matrix σ. (The choice $\alpha = 4$ happens to provide sufficient accuracy for the problem at hand.)

The input to the algorithm is the pair of $K^2 \times K^2$ penalty matrices P_0 and P_1, which effectively specify the matrices Π_0^0, Π_1^0, Π_0^1, and Π_1^1. The algorithm also refers to constant values $\alpha = 4$, $\gamma = 4/3$, $\varepsilon = 1/64$ and $\delta = \varepsilon/\alpha^2$, which, for the sake of clarity, are referred to by these names rather than their numerical values.

1. Set

   $$\rho_0^{(0)} \leftarrow \frac{1}{K^2}\mathbb{1}\otimes\mathbb{1}, \quad \rho_1^{(0)} \leftarrow \frac{1}{K^2}\mathbb{1}\otimes\mathbb{1} \quad \text{and} \quad \sigma^{(0)} \leftarrow \frac{1}{K}\mathbb{1},$$

   for $\mathbb{1}$ denoting the $K \times K$ identity matrix. (These matrices represent completely random quantum states, reflecting the fact that no information is present in the initial guesses for ρ_0, ρ_1, and σ.)
2. Set the maximum iteration number of the algorithm to

   $$T \leftarrow \left\lceil \frac{8\log(K)}{\varepsilon^2\delta} \right\rceil$$

   and repeat the following steps for each t from 0 to T–1:

   (a) Compute the $K \times K$ matrices

   $$M_0^{(t)} \leftarrow \gamma\sigma^{(t)} - \text{PartialTrace}\,(P_0\rho_0^{(t)}P_0)$$

   $$M_1^{(t)} \leftarrow \gamma\sigma^{(t)} - \text{PartialTrace}\,(P_1\rho_1^{(t)}P_1).$$

   (b) Compute the $K \times K$ matrices $\Delta_0^{(t)}$ and $\Delta_1^{(t)}$ that project onto the negative eigenspaces of $M_0^{(t)}$ and $M_1^{(t)}$, respectively.

   (c) Compute

   $$\beta_0^{(t)} \leftarrow \text{Trace}\,((\Delta_0^{(t)}\otimes\mathbb{1})\,P_0\rho_0^{(t)}P_0)$$

   $$\beta_1^{(t)} \leftarrow \text{Trace}\,((\Delta_1^{(t)}\otimes\mathbb{1})\,P_1\rho_1^{(t)}P_1)$$

and set $\beta^{(t)} \leftarrow (\beta_0^{(t)} + \beta_1^{(t)})/2$. If $\beta^{(t)} \le \varepsilon$ then output 1 and stop.

(d) Set

$$X_0^{(t+1)} \leftarrow \exp\left(-\varepsilon\delta \sum_{j=0}^{t} P_0(\Delta_0^{(j)} \otimes \mathbb{1})P_0/\beta^{(j)}\right)$$

$$X_1^{(t+1)} \leftarrow \exp\left(-\varepsilon\delta \sum_{j=0}^{t} P_1(\Delta_1^{(j)} \otimes \mathbb{1})P_1/\beta^{(j)}\right)$$

and set

$$\rho_0^{(t+1)} \leftarrow \frac{2X_0^{(t+1)}}{\text{Trace}\left(X_0^{(t+1)} + X_1^{(t+1)}\right)}$$

$$\rho_1^{(t+1)} \leftarrow \frac{2X_1^{(t+1)}}{\text{Trace}\left(X_0^{(t+1)} + X_1^{(t+1)}\right)}.$$

Also set

$$Y^{(t+1)} \leftarrow \exp\left(\varepsilon\delta \sum_{j=0}^{t} \left(\Delta_0^{(j)} + \Delta_1^{(j)}\right)/\beta^{(j)}\right)$$

and set

$$\sigma^{(t+1)} \leftarrow \frac{Y^{(t+1)}}{\text{Trace}\,(Y^{(t+1)})}.$$

3. If the algorithm has not output 1 and stopped during any iteration of step 2(c), then output 0 and stop.

The correctness of the algorithm is certainly not obvious from its description, and is discussed in the section following this one. In addition to the algorithm's correctness, it is of course also necessary for it to have a highly parallel implementation. As the number of iterations is very small (at most $O(\log K)$, which is polynomially related to the problem size n), it suffices to observe that each iteration can itself be implemented by a highly parallel computation. This is possible through the use of known parallel algorithms for various matrix and algebraic problems.

5. ANALYSIS OF THE ALGORITHM

5.1. Intuition behind the algorithm

The SDP described in the previous section asks for the maximum trace of the average $(Q_0 + Q_1)/2$, over all positive semidefinite matrices Q_0 and Q_1 for which $\sigma - \text{PartialTrace}\,(P_0Q_0P_0)$ and $\sigma - \text{PartialTrace}\,(P_1Q_1P_1)$ are positive semidefinite—for some choice of a density matrix σ. While the formal analysis of the algorithm indeed makes use of the properties of this SDP, the algorithm itself is more naturally viewed as solving a *feasibility problem*.

In particular, for every iteration t the algorithm computes matrices $\rho_0^{(t)}$, $\rho_1^{(t)}$, and $\sigma^{(t)}$ so that $\rho_0^{(t)}$ and $\rho_1^{(t)}$ have average trace equal to 1, but may fail to satisfy the constraint that

$$\sigma^{(t)} - \text{PartialTrace}\,(P_b\rho_b^{(t)}P_b) \qquad (3)$$

is positive semidefinite for $b = 0$, $b = 1$, or both. Successive iterations attempt to bring these matrices closer and closer to satisfying these constraints. If the constraints are "close enough" to being satisfied, the algorithm outputs 1, concluding that the optimal value of the SDP must be close to 1 by taking Q_0 and Q_1 close to $\rho_0^{(t)}$ and $\rho_1^{(t)}$. If the algorithm fails to make sufficient progress toward meeting the constraints, even after T iterations have passed, it outputs 0, concluding that the optimal value of the SDP cannot be close to 1.

Suppose, for a given iteration t, that the algorithm's current choices of $\rho_0^{(t)}$, $\rho_1^{(t)}$, and $\sigma^{(t)}$ are far from satisfying the constraint (3) for $b = 0$, $b = 1$, or both. To be more precise, suppose that one of the matrices

$$\gamma\sigma^{(t)} - \text{PartialTrace}\,(P_b\rho_b^{(t)}P_b)$$

fails to be positive semidefinite (for $\gamma = 4/3$). In this case, the subspace corresponding to the projection $\Delta_b^{(t)}$ onto the negative part of this matrix represents a part of the search space where $\rho_0^{(t)}$, $\rho_1^{(t)}$ and $\sigma^{(t)}$ are inadequate: if progress toward meeting the constraints is to be made, σ must be made larger on this subspace while $\text{PartialTrace}(P_b\rho_bP_b)$ must be smaller.

In both cases, this modification is made by including an additional term in the argument to the matrix exponential that is normalized to produce the next iteration's choices $\rho_0^{(t+1)}$, $\rho_1^{(t+1)}$, and $\sigma^{(t+1)}$. This has the effect of shifting the weight of $\rho_0^{(t+1)}$ and $\rho_1^{(t+1)}$ away from the parts of $\rho_0^{(t)}$, and $\rho_1^{(t)}$ that contributed to the constraint violations indicated by $\Delta_0^{(t)}$, and $\Delta_1^{(t)}$, respectively, and likewise shifting the weight of $\sigma^{(t+1)}$ toward these parts of $\sigma^{(t)}$. (A negative term in the exponential reduces weight while a positive term increases weight.)

5.2. Comments on the formal algorithm analysis

That the algorithm works correctly is, of course, not as simple as the intuitive discussion above might suggest, due in large part to the complicated behavior of the exponential of a sum of matrices. A proper analysis of the algorithm is necessarily technical, and one can be found in the full version of this paper. This extended abstract will not present a detailed formal analysis, but will instead present just a high-level sketch of the analysis, intended only to provide the reader with its basic ideas.

A rigorous analysis of the algorithm makes use of the notion of semidefinite programming duality. Every semidefinite programming problem, stated as a maximization problem typically called the *primal problem*, has a corresponding *dual problem*, which is a minimization problem. The methodology for determining the dual problem corresponding to a given primal problem is standard, but will not be discussed here. Each possible value of the dual problem provides an upper bound on the optimal value of the primal problem.

For the particular SDP at hand, the dual problem is to minimize the largest eigenvalue of the average $\frac{1}{2}R_0 + \frac{1}{2}R_1$, over all $K \times K$ positive semidefinite matrices R_0 and R_1 obeying the constraint that every eigenvalue of $P_0(R_0 \otimes \mathbb{1})P_0$ and $P_1(R_1 \otimes \mathbb{1})P_1$ is at least 1, or equivalently that the matrices

$$P_0(R_0 \otimes \mathbb{1})P_0 - \mathbb{1} \otimes \mathbb{1} \quad \text{and} \quad P_1(R_1 \otimes \mathbb{1})P_1 - \mathbb{1} \otimes \mathbb{1}$$

are both positive semidefinite. (As before, $\mathbb{1}$ denotes the $K \times K$ identity matrix.)

The analysis works by constructing solutions to either the primal problem or dual problem based on the behavior of the algorithm, guaranteeing its correctness under the

assumption that the optimal value of the primal problem is either very close to 1 or to 1/2.

The simpler case is that the algorithm outputs 1. This must happen during some particular iteration t, and for this choice of t one may consider the matrices

$$Q_0 = \frac{\rho_0^{(t)}}{\gamma + 4\beta^{(t)}} \quad \text{and} \quad Q_1 = \frac{\rho_1^{(t)}}{\gamma + 4\beta^{(t)}}.$$

For an appropriate choice of σ, which can be constructed from $\rho_0^{(t)}$, $\rho_1^{(t)}$, $\sigma^{(t)}$, $\Delta_0^{(t)}$, $\Delta_1^{(t)}$ and $\beta^{(t)}$, it holds that both σ – PartialTrace $(P_0 Q_0 P_0)$ and σ – PartialTrace $(P_1 Q_1 P_1)$ are positive semidefinite. The trace of the average of Q_0 and Q_1 is at least $1/(\gamma + 4\varepsilon) > 5/8$. It follows that the optimal value of the primal problem cannot be too close to 1/2, so it is necessarily close to 1.

The more difficult case is that the algorithm outputs 0. In this case, the matrices

$$R_0 = \frac{1+4\varepsilon}{T}\sum_{t=0}^{T-1} \Delta_0^{(t)}/\beta^{(t)} \quad \text{and} \quad R_1 = \frac{1+4\varepsilon}{T}\sum_{t=0}^{T=1} \Delta_1^{(t)}/\beta^{(t)}$$

form a solution to the dual problem satisfying the required constraints and achieving a value smaller than 7/8. Establishing that these matrices indeed satisfy the required constraints and achieve a value smaller than 7/8 is somewhat technical, but follows from a basic methodology that appeared in previous works (including Warmuth and Kuzmin[16] and Arora and Kale[1]). With a value of at most 7/8 for the dual problem, the same value has been established as an upper-bound for the primal problem. It follows that the optimal value for the primal problem is sufficiently far from 1 that it must be close to 1/2.

There is a small complication in the formal analysis that does not present a major obstacle, but is nevertheless worthy of note. The complication is that the algorithm cannot perform all of the required computations exactly, but must approximate some of them. (In particular, the matrix exponentials and the negative eigenspace projections can only be approximated, albeit with very high accuracy.) A slight modification of the analysis sketched above demonstrates, however, that the algorithm is not particularly sensitive to errors. The steps of the computation may, in fact, be performed with significantly less accuracy than is possible within the resource constraints required of the algorithm.

6. CONCLUSION

The characterization QIP = PSPACE implies that quantum computation does not provide an increase in the expressive power of interactive proof systems. It is tempting to view this fact as a negative result for quantum computing, but this view is not well justified. What is most interesting about quantum computation is its potential in a standard computational setting, where an algorithm (deterministic, probabilistic, or quantum) receives an input and produces an output in isolation, as opposed to through an interaction with a hypothetical prover. The main result of this paper has no implications to this fundamental question. A more defensible explanation for the equivalence of quantum and classical computing in the interactive proof system model is the model's vast computational power: all of PSPACE. That such power washes away the difference between quantum and classical computing is, in retrospect, perhaps not unexpected.

One future research direction that is clearly suggested by this paper is: what class of SDPs can be solved approximately by parallel algorithms? We do not have an answer to this question, and believe it is certainly deserving of further investigation. There are, in addition, still many open questions relating to variants of quantum interactive proof systems—the most notable being the *multiprover quantum interactive proof system* model. This model is, in fact, currently so poorly understood that it is not even known if every problem having a multiprover quantum interactive proof system is necessarily decidable. C

References

1. Arora, S., Kale, S. A combinatorial, primal-dual approach to semidefinite programs. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing* (2007), 227–236.
2. Babai, L. Trading group theory for randomness. In *Proceedings of the 17th Annual ACM Symposium on Theory of Computing* (1985), 421–429.
3. Bennett, C., Brassard, G. Quantum cryptography: Public key distribution and coin tossing. In *Proceedings of the IEEE International Conference on Computers, Systems, and Signal Processing* (1984), 175–179.
4. Bennett, C., Brassard, G., Crépeau, C., Jozsa, R., Peres, A., Wootters, W. Teleporting an unknown quantum state via dual classical and EPR channels. *Phys. Rev. Lett. 70*, 12 (1993), 1895–1899.
5. Borodin, A. On relating time and space to size and depth. *SIAM J. Comput. 6* (1977), 733–744.
6. Deutsch, D. Quantum theory, the Church–Turing principle and the universal quantum computer. *Proc. R.Soc. Lond. A400* (1985), 97–117.
7. Feynman, R. Simulating physics with computers. *Int. J. Theor. Phys. 21*, 6/7 (1982), 467–488.
8. Goldwasser, S., Micali, S., Rackoff, C. The knowledge complexity of interactive proof systems. *SIAM J. Comput., 18*, 1 (1989), 186–208. A preliminary version appeared in *Proceedings of the 17th Annual ACM Symposium on Theory of Computing* (1985), 291–304.
9. Kitaev, A., Watrous, J. Parallelization, amplification, and exponential time simulation of quantum interactive proof system. In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing* (2000), 608–617.
10. Lund, C., Fortnow, L., Karloff, H., Nisan, N. Algebraic methods for interactive proof systems. *J. ACM 39*, 4 (1992), 859–868. A preliminary version appeared in *Proceedings of the 31st Annual IEEE Symposium on Foundations of Computer Science* (1990), 2–10.
11. Marriott, C., Watrous, J. Quantum Arthur-Merlin games. *Comput. Complex. 14*, 2 (2005), 122–152.
12. Nielsen, M., Chuang, I. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.
13. Raz, R. Exponential separation of quantum and classical communication complexity. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing* (1999), 358–376.
14. Shamir, A. IP = PSPACE. *J. ACM 39*, 4 (1992), 869–877. A preliminary version appeared in *Proceedings of the 31st Annual IEEE Symposium on Foundations of Computer Science* (1990), 11–15.
15. Shor, P. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM J. Comput. 26*, 5 (1997), 1484–1509. A preliminary version appeared with the title "Algorithms for quantum computation: discrete logarithms and factoring" in *Proceedings of the 35th Annual IEEE Symposium on Foundations of Computer Science* (1994), 124–134.
16. Warmuth, M., Kuzmin, D. Online variance minimization. In *Proceedings of the 19th Annual Conference on Learning Theory*, volume 4005 of *Lecture Notes in Computer Science* (Springer, 2006), 514–528.
17. Watrous, J. PSPACE has constant-round quantum interactive proof systems. In *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science* (1999), 112–119.

Rahul Jain, Department of Computer Science and Centre for Quantum Technologies, National University of Singapore, Republic of Singapore.

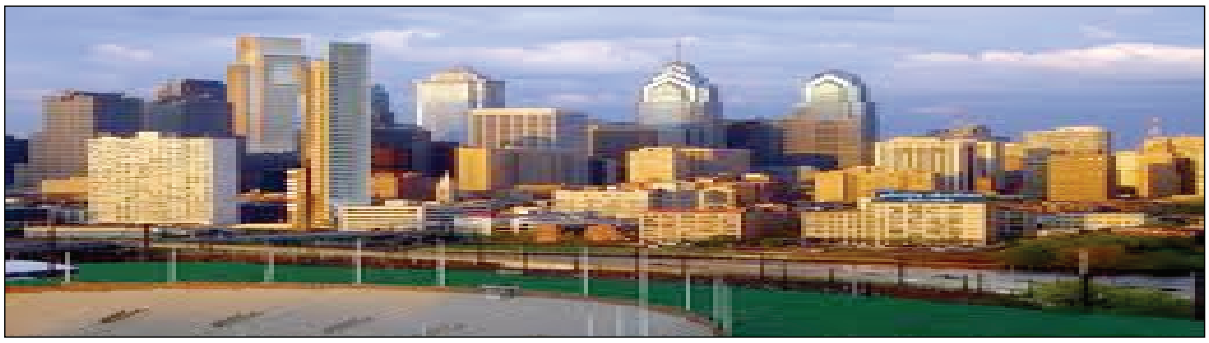
Zhengfeng Ji, Perimeter Institute for Theoretical Physics, Waterloo, Ontario, Canada.

Sarvagya Upadhyay and John Watrous, Institute for Quantum Computing and School of Computer Science, University of Waterloo Waterloo, Ontario, Canada.

CALL FOR PARTICIPATION

CTS 2011

Philadelphia, Pennsylvania, USA



The 2011 International Conference on Collaboration Technologies and Systems

May 23 – 27, 2011
The Sheraton University City Hotel
Philadelphia, Pennsylvania, USA

Important Dates:

Paper Submission Deadline	**December 15, 2010**
Workshop/Special Session Proposal Deadline	**November 15, 2010**
Tutorial/Demo/Panel Proposal Deadline	**January 8, 2011**
Notification of Acceptance	**February 1, 2011**
Final Papers Due	**March 1, 2011**

For more information, visit the CTS 2011 web site at:
http://cts2011.cisedu.info/

In cooperation with the ACM, IEEE, IFIP

Ada Core Technologies, Inc.

Sr. Software Engineer
(New York, NY)

Design, develop and test software modifications and enhancements to GNAT Pro tool chain. Enhance, debug, support, maintain and test AdaCore products on diverse development platforms, hardware and technologies. Support, maintain and document software functionality. Develop test programs for GNAT Pro and related tools. Send resumes to Edmond Schonberg, VP, Ada Core Technologies, Inc, 104 Fifth Ave, 15th Fl, New York, NY 10011. No calls, faxes or emails please! EOE.

Air Force Institute of Technology (AFIT) Dayton, Ohio

Department of Electrical and Computer Engineering
Graduate School of Engineering and Management
Faculty Positions in Computer Science or Computer Engineering

The Department of Electrical and Computer Engineering is seeking applicants for tenure track positions in computer science or computer engineering. The department is particularly interested in receiving applications from individuals with strong backgrounds in formal methods (with emphasis on cryptography), software engineering, bioinformatics, computer architecture/VLSI systems, and computer networks and security. The positions are at the assistant professor level, although qualified candidates will be considered at all levels. Applicants must have an earned doctorate in computer science or computer engineering or closely related field and must be U.S. citizens. These positions require teaching at the graduate level as well as establishing and sustaining a strong research program.

AFIT is the premier institution for defense-related graduate education in science, engineering, advanced technology, and management for the U.S. Air Force and the Department of Defense (DoD). Full details on these positions, the department, and application procedures can be found at: http://www.afit.edu/en/eng/employment_faculty.cfm

Review of applications will begin immediately and will continue until the positions are filled. The United States Air Force is an equal opportunity, affirmative action employer.

American University of Nigeria

Dean School of Information Technology and Communications

The American University of Nigeria (http://www.aun.edu.ng/), in partnership with **The American University, Washington, D.C.,** seeks qualified candidates for **Dean of the School of ITC.** U.S. PhD preferred; five to seven years of experience; management abilities; and effective organizational, and communication skills. Applicants with extensive international experience are encouraged to apply. Qualified applicants should submit a cover letter and CV to: AUNDeanITC@HCALtd.com.

AUN, a private university, based in Yola, the capital of Adamawa State in Northeast Nigeria, offers a BS in Computer Science, Software Engineering, Information Systems, and Communications.

Arizona State University

Post Doctoral Scholar / Research Scientist

Immediate two-year (renewable) position: NSF-Funded Affective Meta Tutor project advancing Intelligent Tutoring Systems and Affective Learning Companions for dynamic systems modeling. See: http://www.public.asu.edu/~kvanlehn/ and hci.asu.edu

Baylor University

Chairperson, Department of Computer Science

Chartered in 1845 by the Republic of Texas, Baylor University is the oldest university in Texas and the world's largest Baptist University. Baylor's mission is to educate men and women for worldwide leadership and service by integrating academic excellence and Christian commitment within a caring community. Baylor is actively recruiting new faculty with a strong commitment to the classroom and an equally strong commitment to discovering new knowledge as Baylor aspires to become a top tier research university while reaffirming and strengthening its distinctive Christian mission as described in Baylor 2012 (www.baylor.edu/vision/).

The Department of Computer Science seeks qualified candidates for the faculty chair position. Viable candidates possess high academic standards and Christian convictions matched by an active participation in a community of faith. Additionally, s/he must have a Ph.D. in Computer Science or related area, leadership experience, a commitment to undergraduate and graduate education, a strong research record that includes significant external funding, effective communication and organization skills.

The Department: The Department offers a CSAB-accredited B.S. in Computer Science degree, a B.A. degree with a major in Computer Science, a B.S. in Informatics with a major in Bioinformatics, and a M.S. degree in Computer Science. The Department has 15 full-time faculty, over 370 undergraduate majors and 30 master's students. The Department's greatest strength is its dedication to the success of the students and each other. Interested candidates may contact any faculty member to ask questions and/or visit the web site of the School of Engineering and Computer Science at http://www.ecs.baylor.edu.

The University: Baylor University, situated on a 500-acre campus next to the Brazos River. It annually enrolls more than 14,000 students in over 150 baccalaureate and 80 graduate programs through: the College of Arts and Sciences; the Schools of Business, Education, Engineering and Computer Science, Music, Nursing, Law, Social Work, and Graduate Studies; plus Truett Seminary and the Honors College. For more information see http://www.baylor.edu.

Application Procedure: Applications, including detailed curriculum vitae, a statement demonstrating an active Christian faith, and contact information for three references should be sent to: Chair Search Committee, Department of Computer Science, Baylor University, One Bear Place #97356, Waco, TX 76798-7356.

Appointment Date: Fall 2011. For full consideration, applications should be received by January 1, 2011. However, applications will be accepted until the position is filled.

Baylor is a Baptist university affiliated with the Baptist General Convention of Texas. As an Affirmative Action/Equal Employment Opportunity employer, Baylor encourages minorities, women, veterans, and persons with disabilities to apply.

Boston University

Department of Electrical & Computer Engineering
Faculty Search

The Department of Electrical & Computer Engineering (ECE) at Boston University (BU) is seeking candidates for several anticipated faculty positions. While all areas and ranks are under consideration, there is particular interest in entry-level and mid-career candidates in the various areas of Computer Engineering, including software, security, embedded systems, and hardware design. The Department is seeking to foster growth in the broad, interdisciplinary topics of energy, health, and communications; candidates with research interests that transcend the traditional boundaries of ECE are strongly encouraged to apply. Joint appointments with other BU departments and with the Divisions of Material Science & Engineering and Systems Engineering are possible for candidates with appropriate interests.

Qualified candidates must possess a relevant, earned PhD, and have a demonstrable ability to teach effectively, develop funded research programs in their area of expertise, and contribute to the tradition of excellence in research that is characteristic of the ECE Department. Self-motivated individuals who thrive on challenge and are eager to utilize their expertise to strengthen an ambitious program of departmental enhancement are desired. Women, minorities, and candidates from other underrepresented groups are especially encouraged to apply and help us continue building an exceptional 21st century university department.

ECE at BU is a world-class, progressive department with excellent resources that is steadily gaining national and international prominence for its exceptional research and education record. ECE is part of BU's rapidly growing and innovative College of Engineering, and currently consists of 44 faculty members, 200 graduate students, and 250 BS majors. Outstanding collaboration opportunities are available with nationally recognized medical centers and universities/colleges, nearby research centers, and industry throughout the Boston area.

Beyond its research and academic activities, BU has a lively, urban campus situated along the banks of the Charles River in Boston's historic Fenway-Kenmore neighborhood. The campus and surrounding areas offer limitless opportunities for recreational activities, from world-class art and performances to sporting events and fine dining.

Please visit http://www.bu.edu/ece/faculty-search for instructions on how to apply. Application deadline is January 31, 2011. Boston University is an Equal Opportunity/Affirmative Action Employer.

Cal Poly Pomona
Assistant Professor

The Computer Science Department invites applications for a tenure-track position at the rank of Assistant Professor to begin Fall 2011. We are particularly interested in candidates with specialization in Software Engineering, although candidates in all areas of Computer Science will be considered, and are encouraged to apply. Cal Poly Pomona is 30 miles east of L.A. and is one of 23 campuses in the California State University. The department offers an ABET-accredited B.S. program and an M.S. program. Qualifications: Possess, or complete by September 2011, a Ph.D. in Computer Science or closely related area. Demonstrate strong English communication skills, a commitment to actively engage in the teaching, research and curricular development activities of the department at both undergraduate and graduate levels, and ability to work with a diverse student body and multicultural constituencies. Ability to teach a broad range of courses, and to articulate complex subject matter to students at all educational levels. First consideration will be given to completed applications received no later than December 15, 2010. Contact: Faculty Search Committee, Computer Science Department, Cal Poly Pomona, Pomona, CA 91768. Email: cs@csupomona.edu. Cal Poly Pomona is an Equal Opportunity, Affirmative Action Employer. Position announcement available at: http://academic.csupomona.edu/faculty/positions.aspx. Lawful authorization to work in US required for hiring.

California University of PA
Computer Science Faculty

California University of Pennsylvania invites applications for this tenure-track faculty position. Ability to effectively teach Computer Science courses and a Ph. D. in Computer Science is requires. For position details and to apply, visit https://careers.cup.edu. Calu is M/F/V/D/AA/EOE.

Carnegie Mellon University
School of Design
Tenure Track Faculty Position

School of Design at Carnegie Mellon University
Tenure-Track Faculty Position –
Design and Computation
Position begins August 2011.
Apply to kh@cmu.edu.
View complete job description at
http://bit.ly/cjwnFs

Carnegie Mellon University
Tepper School of Business
Information Systems Tenure-Track Junior Faculty Position

Tenure-track faculty opening in Information Systems at the Assistant Professor level, starting in September 2011. *We are looking for people who are interested in how IT may transform businesses, markets, and economic processes.* We have a particular interest in applicants with research and teaching interests in the development and application of machine learning techniques to business problems or in other related areas of business analytics. Teaching assignments encompass BS, Masters, and Ph.D. programs. Applicants should send a current curriculum vita, evidence of research such as publications, working papers, or dissertation proposal to: isgroup@andrew.cmu.edu and three recommendation letters (*via the Postal Service*) to Mr. Phillip Conley, Information Systems Faculty Recruiting, Carnegie Mellon University, Tepper School of Business, Room

UNC CHARLOTTE
College of Computing and Informatics

MULTIPLE FACULTY POSITIONS

College of Computing and Informatics, University of North Carolina at Charlotte

CCI is one of the few college-level organizations in computing and informatics among major research universities in the US, with close to 60 faculty, 30 staff, and 1200 students, including 130 Ph.D. students. The college has a vibrant and cutting edge research enterprise, most notably: Charlotte Visualization Center is one of the leading centers in the nation sponsored by the Department of Homeland Security in visual analytics; Cyber DNA Center is a National Center of Excellence in Information Assurance Research and Education designated by the National Security Agency; the Bioinformatics Research Center leads the University in structural bioinformatics, molecular biophysics, plant genomics, and metagenomics research. For fiscal year 2009-2010 the College received over $15 million in external research funding.

UNC Charlotte is a rapidly growing urban research university with 900 faculty, over 25,000 students, including 5,400 graduate students, with a projected enrollment of 35,000 by 2020. It is ranked among the top 10 of up-and-coming national universities by the US News and World Report. Charlotte is a dynamic and diverse region of 1.8 million people and is one of the most livable major urban areas in the nation, with outstanding cultural and recreational amenities. It is the second largest banking center in the country, and a leading center for energy, healthcare, retail, and logistics industries.

The College of Computing and Informatics is embarking on an exciting venture to develop a new class of leading computing and informatics programs for the 21st Century talent and innovation needs. Our strategy emphasizes the interplay between developing critical mass in analytics, security, and informatics, and building strong connections with banking, healthcare, energy, life science, and biotechnology. We invite outstanding entrepreneurial, thought leaders to join our faculty in the following positions:

1. **Department of Computer Science:** tenure-track full professor. The successful candidate should have demonstrated skills in working across college and departmental boundaries and is experienced in building and managing large-scale, interdisciplinary research and educational efforts. The candidate's research can lie in serious games, data warehousing, data analytics, or knowledge systems and discovery.
2. **Department of Software and Information Systems:** two tenure-track faculty positions at all levels, with strong preference given to full and associate professor ranks. A successful candidate must have an excellent research record that can attract substantial research funding. The Department is particularly interested in faculty with research expertise in health informatics, information and network security, modeling, and simulation of complex systems.
3. **Department of Bioinformatics and Genomics** invites applications for the Carol Grotnes Belk Distinguished Professorship of Bioinformatics and Genomics with tenure.

Salaries for the above positions will be highly competitive. All candidates must have a Ph.D. degree in relevant areas. For application details, please visit (https://jobs.uncc.edu) and click on faculty. Review of applications will start in September 2010 and continue until positions are filled. The University of North Carolina at Charlotte is an EOE/AA employer and an NSF ADVANCE Institution.

369 Posner Hall, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890 (Phone: 412-268-6212). **In order to ensure full consideration, completed applications must be received by Friday, JANUARY 14, 2011.**

Applicants may hold a doctoral degree in any business discipline, Information Systems, Computer Science, Economics or Operations Research. We are primarily seeking candidates at the Assistant Professor level. Applicants should have completed or be nearing completion of a Ph.D., and should demonstrate potential of excellence in research and teaching.

For information on Carnegie Mellon University's Tepper School of Business, please visit www.tepper.cmu.edu

Carnegie Mellon is an equal opportunity, affirmative action employer with particular interest in identifying women and minority applicants for faculty positions.

Central Michigan University
Department of Computer Science

The department has two tenure track positions in Information Technology to be filled in Fall 2011. One is in Applied Networking and one is in Medical or Health Informatics. For more information and to apply electronically: www.jobs.cmich.edu.

Central Washington University
Assistant/Associate Professor

Central Washington University, Computer Science Dept - accepting applications for Ass't/Assoc Prof. All specialization areas are welcome. To apply online, visit: https://jobs.cwu.edu Screening date: 1/03/2011 AA/EEO/Title IX Institution.

Eastern Washington University
Assistant Professor

The Computer Science Department at Eastern Washington University invites applications for a tenure-track position starting Sept 2011. Please visit: http://access.ewu.edu/HRRR/Jobs.xml for complete information. For questions contact Margo Stanzak (509) 359-4734

Florida International University
Tenure-track & Tenured Faculty Positions (All levels)

FIU is a multi-campus public research university located in Miami, a vibrant and globally connected 24/7 city. Miami's captivating skyline, tasteful tropical cuisine, vivid arts, historically rich and diverse neighborhoods, trendy South Beach scene, bustling international trade, and youthful exuberance provide a perfect environment for our engaged university.

Serving more than 42,000 students, FIU offers more than 180 baccalaureate, masters, professional, and doctoral degree programs. As one of South Florida's anchor institutions, FIU is worlds ahead in its local and global engagement, finding solutions to the most challenging problems of our times.

The School of Computing and Information Sciences seeks exceptionally qualified candidates for **tenure-track and tenured faculty positions at all levels** and in all areas of Computer Science. A Ph.D. in Computer Science or related disciplines is required. Preference will be given to candidates who will enhance or complement our existing research strengths in data and information management, informatics, operating systems, networking, programming languages, security, and software engineering. Ideal candidates for junior positions should have a record of exceptional research in their early careers. Candidates for senior positions must have an active and proven record of excellence in funded research, publications, and professional service, as well as a demonstrated ability to develop and lead collaborative research projects. In addition to developing or expanding a high-quality research program, all successful applicants must be committed to excellence in teaching at both graduate and undergraduate levels.

Florida International University (FIU), the state university of Florida in Miami, is ranked by the Carnegie Foundation as a comprehensive doctoral research university with high research activity. The School of Computing and Information Sciences (SCIS) is a rapidly growing program of excellence at the University, with 32 faculty members and over 1,400 students, including over 90 Ph.D. students. SCIS offers B.S., M.S., and Ph.D. degrees in Computer Science, an M.S. degree in Telecommunications and Networking, and B.S., B.A., and M.S. degrees in Information Technology. SCIS has received approximately

$8.4M in the last two years in external research funding, has six research centers/clusters with first-class computing infrastructure and support, and enjoys broad and dynamic industry and international partnerships.

HOW TO APPLY:
Applications, including a letter of interest, contact information, curriculum vitae, and the names of at least three references, should be submitted directly to the FIU J.O.B.S Link website at https://www.fiujobs.org; refer to Position # 45534. The application review process will begin on January 17, 2011, and will continue until the position is filled. Further information can be obtained from the School website http://www.cis.fiu.edu, or by e-mail to recruit@cis.fiu.edu.

FIU is a member of the State University System of Florida and is an Equal Opportunity, Equal Access Affirmative Action Employer.

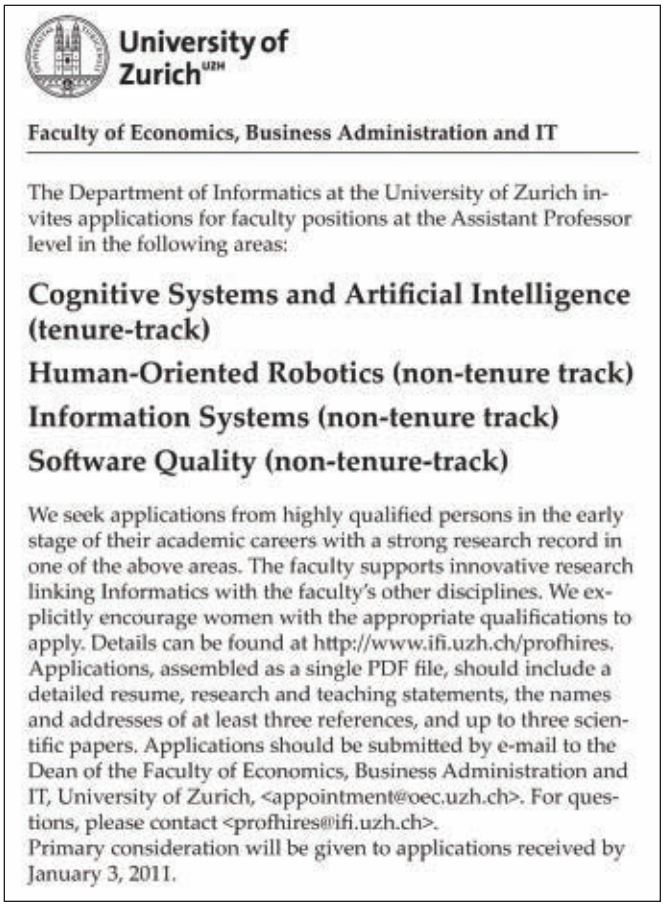
Georgia State University
Department of Computer Science

The Department of Computer Science of Georgia State University invites applications for an anticipated position of Full Professor (possibly Eminent Scholar) in the bioinformatics area beginning the Fall semester, 2011, pending budgetary approval. Earned Ph.D. in Computer Science or a related discipline is required. An offer of employment will be conditional on background verification.

Prospective candidates should demonstrate ability to bring national and international recognition to the department as a center of excellence for bioinformatics research and education. The hired eminent scholar is expected to bring in major extramural funding, mentor junior faculty, recruit top quality PhD students, and foster interdisciplinary collaborations amongst faculty in various departments in GSU.

Georgia State University, founded in 1913, is a Carnegie Doctoral/Research Extensive university. Located in the heart of downtown Atlanta, this major research university has an enrollment of more than 30,000 undergraduate and graduate students in six colleges. Georgia State is the second largest university in the state, with students coming from every county in Georgia, every state in the nation and from over 145 countries. Georgia State University is currently embarking on a record $1 billion campus expansion. The Computer Science Department offers programs leading to the B.S., M.S., and Ph.D. degrees in computer science. Currently, 20 out of more than 60 Ph.D. students are involved in bioinformatics research. They are supervised by 10 faculty members fully or substantially involved in bioinformatics research through collaboration with Computer Science and Biology faculty. Departmental computing facilities for research and instruction include a departmental network of PCs, Unix/Linux workstations, two interconnected Beowulf clusters, and a 24-processor supercomputer. The department's faculty attracts substantial from many federal agencies including five NSF CAREER Awards.

Applicants should send letter of interest, C.V., and three letters of recommendation to:



Dr. Yi Pan, Chair
Department of Computer Science
Georgia State University
34 Peachtree Street, Suite 1450
Atlanta, Georgia, 30303

Applications can also be sent via email to pan@cs.gsu.edu and will be accepted until position is filled.

Georgia State University, a Research University of the University System of Georgia, is an AA/EEO employer.

International Computer Science Institute

Director

The International Computer Science Institute (ICSI), an independent non-profit laboratory closely affiliated with the EECS Department, University of California, Berkeley (UCB), invites applications for the position of Director, beginning Fall 2011.

The ICSI Director's primary responsibilities are to: oversee and expand ICSI's research agenda; act as a high-level external evangelist for ICSI research; identify and pursue strategic funding opportunities; and strengthen ICSI's relationship with UCB. The Director reports directly to ICSI's Board of Trustees.

ICSI is recognized for world-class research activities in networking, speech, language and vision processing, as well as computational biology and computer architecture. Several of ICSI's research staff have joint UCB appointments, and many UCB graduate students perform their research at ICSI. In addition, ICSI places significant emphasis on international partnerships and visiting scholar programs.

ICSI is seeking a Director with sufficient breadth, interest, and professional connections to promote and augment ICSI's ongoing research efforts. Applicants should have recognized research leadership, as well as a strong record in research management and demonstrated success at government and industrial fundraising. Experience with international collaboration and fundraising is a plus.

Applications should include a resume, selected publications, and names of three references. Review begins February 1, 2011; candidates are urged to apply by that date.

To learn more about ICSI, go to http://www.icsi.berkeley.edu.

To apply for this position, send the above material to apply@icsi.berkeley.edu. Recommenders should send letters directly to apply@icsi.berkeley.edu by 2/1/2011. ICSI is an Affirmative Action/Equal Opportunity Employer. Applications from women and minorities are especially encouraged.

Loyola University Maryland

Assistant Professor, Computer Science

The Computer Science Department invites applications for a Assistant Professor position for the academic year 2011-2012. We are seeking enthusiastic individuals, committed to excellent teaching and continued scholarship. A Ph.D. in Computer Science, Computer Engineering, or a closely related field is required. While candidates in all areas of specialization will be considered, we especially welcome applicants with expertise in software engineering or network security. More information is available at www.loyola.edu/cs and www.loyola.edu

Applicants must submit the following online (http://careers.loyola.edu): a letter of application, curriculum vita, outline of teaching philosophy, and a statement of research objectives.

Massachusetts Institute of Technology

Faculty Positions

The Department of Electrical Engineering and Computer Science (EECS) seeks candidates for faculty positions starting in September 2011. Appointment would be at the assistant or untenured associate professor level. In special cases, a senior faculty appointment may be possible. Faculty duties include teaching at the graduate and undergraduate levels, research, and supervision of student research. We will consider candidates with backgrounds and interests in any area of electrical engineering and computer science. Faculty appointments will commence after completion of a doctoral degree.

Candidates must register with the EECS search website at

https://eecs-search.eecs.mit.edu, and must submit application materials electronically to

Virginia Tech — Invent the Future

Tenure-Track Assistant Professor Position in Compilers
Department of Computer Science, Virginia Tech

The Department of Computer Science at Virginia Tech (www.cs.vt.edu) invites applications for hiring at the Assistant Professor rank from candidates with primary research focus in compilers, including but not limited to (a) compilers that support programming languages and models for emerging architectures such as many-core processors and GPGPUs, (b) program analysis, (c) optimizing and parallelizing compilers, and (d) compiler-driven runtime systems. The position is part of a cluster hire of a total of five positions on both the Blacksburg and National Capital Region campuses of Virginia Tech in computer systems, security, and cybersecurity by the Department of Computer Science and the Bradley Department of Electrical and Computer Engineering (ECE). The successful candidate will join the department on the Blacksburg campus.

Candidates should have a doctoral degree in Computer Science or a cognate area, a record of significant research achievement and publication, a coherent research and teaching plan showing the potential to secure research funding, build a research program in their area of specialty, and contribute to the department's graduate/undergraduate teaching mission in compilers and related areas, and sensitivity to issues of diversity in the campus community.

Salary for suitably qualified applicants is competitive and commensurate with experience. Virginia Tech is an Equal Opportunity/Affirmative Action Institution. **Applications must be submitted online to https://jobs.vt.edu for posting #0100838.** Applicant screening will begin December 15, 2010 and continue until the position is filled. Inquiries should be directed to Godmar Back, Hiring Committee Chair, gback@cs.vt.edu.

Cybersecurity Senior Position
Department of Computer Science, Virginia Tech

The Department of Computer Science (CS@VT) at Virginia Tech seeks applicants for a tenure-track faculty position in the area of cybersecurity, at Associate Professor or Professor rank, located on the campus of Virginia Tech in the National Capital Region (NCR, www.ncr.vt.edu). The successful candidate will contribute to the research and graduate programs in the NCR and collaborate with faculty at Virginia Tech's campus in Blacksburg, VA. This position is part of a cluster hire of a total of five positions on both campuses in computer systems, security, and cybersecurity by the Department of Computer Science and the Bradley Department of Electrical and Computer Engineering (ECE) at Virginia Tech.

Candidates should have research interests in information security, trustworthy systems, and other topics relevant to national security and national critical infrastructure. Ideal candidates combine cybersecurity with existing department strengths in software engineering, high performance computing, systems and networks, data mining, and human-computer interaction. Candidates should have a record appropriate to a tenured position in scholarship, leadership, and interdisciplinary collaboration in cybersecurity; demonstrated ability to contribute to teaching at the graduate level in cybersecurity and related subjects; sensitivity to issues of diversity in the campus community; established professional network and experience in working with cybersecurity-related government agencies and industry, potentially including classified research, and the skills needed to establish and grow a geographically distributed research group.

Salary for suitably qualified applicants is competitive and commensurate with experience.

Applications must be submitted online to https://jobs.vt.edu for posting #0100818. Applicant screening will begin December 15, 2010 and continue until the position is filled. Inquiries should be directed to Chris North, Hiring Committee Chair, north@cs.vt.edu.

Machine Learning/Artificial Intelligence
Department of Computer Science, Virginia Tech

The Department of Computer Science at Virginia Tech (www.cs.vt.edu) invites applications from candidates in artificial intelligence with particular interests in machine learning for a full-time tenure-track position at any rank, with preference for hiring at the Professor or Associate Professor rank. The department plans on making multiple hires over multiple years in this area. Candidates should have an established record appropriate to the desired rank of scholarship, leadership, and collaboration in a variety of computing and interdisciplinary areas; demonstrated ability to contribute to our department's teaching mission at the undergraduate and graduate levels in AI and related subjects; and the organizational skills needed to establish and grow a multidisciplinary research center.

Salary for suitably qualified applicants is competitive and commensurate with experience.

Applications must be submitted online to https://jobs.vt.edu for posting #0100774. Applicant screening will begin December 15, 2010 and continue until the position is filled. Inquiries should be directed to Dennis Kafura, kafura@cs.vt.edu.

Virginia Tech is an Equal Opportunity/Affirmative Action Institution.

this website. Candidate applications should include a description of professional interests and goals in both teaching and research. Each application should include a curriculum vita and the names and addresses of three or more individuals who will provide letters of recommendation. Candidates should request that their letter writers submit recommendation letters directly to MIT on the website above. Please submit complete application by December 15, 2010.

Send all materials not submitted on the website to:

Professor W. Eric L. Grimson
Department Head, Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Room 38-401
77 Massachusetts Avenue
Cambridge, MA 02139

M.I.T. is an equal opportunity/affirmative action employer.

Max Planck Institute for Software Systems (MPI-SWS)

Tenure-track openings

Applications are invited for tenure-track and tenured faculty positions in all areas related to the study, design, and engineering of software systems. These areas include, but are not limited to, data and information management, programming systems, software verification, parallel, distributed and networked systems, and embedded systems, as well as cross-cutting areas like security, machine learning, usability, and social aspects of software systems. A doctoral degree in computer science or related areas and an outstanding research record are required. Successful candidates are expected to build a team and pursue a highly visible research agenda, both independently and in collaboration with other groups. Senior candidates must have demonstrated leadership abilities and recognized international stature.

MPI-SWS, founded in 2005, is part of a network of eighty Max Planck Institutes, Germany's premier basic research facilities. MPIs have an established record of world-class, foundational research in the fields of medicine, biology, chemistry, physics, technology and humanities. Since 1948, MPI researchers have won 17 Nobel prizes. MPI-SWS aspires to meet the highest standards of excellence and international recognition with its research in software systems.

To this end, the institute offers a unique environment that combines the best aspects of a university department and a research laboratory:

a) Faculty receive generous base funding to build and lead a team of graduate students and post-docs. They have full academic freedom and publish their research results freely.

b) Faculty supervise doctoral theses, and have the opportunity to teach graduate and undergraduate courses.

c) Faculty are provided with outstanding technical and administrative support facilities as well as internationally competitive compensation packages.

MPI-SWS currently has 8 tenured and tenure-track faculty, and is funded to support 17 faculty and about 100 doctoral and post-doctoral positions. Additional growth through outside funding is possible. We maintain an open, international and diverse work environment and seek applications from outstanding researchers regardless of national origin or citizenship. The working language is English; knowledge of the German language is not required for a successful career at the institute.

The institute is located in Kaiserslautern and Saarbruecken, in the tri-border area of Germany, France and Luxembourg. The area offers a high standard of living, beautiful surroundings and easy access to major metropolitan areas in the center of Europe, as well as a stimulating, competitive and collaborative work environment. In immediate proximity are the MPI for Informatics, Saarland University, the Technical University of Kaiserslautern, the German Center for Artificial Intelligence (DFKI), and the Fraunhofer Institutes for Experimental Software Engineering and for Industrial Mathematics.

Qualified candidates should apply online at http://www.mpi-sws.org/application. The review of applications will begin on January 3, 2011, and applicants are strongly encouraged to apply by that date; however, applications will continue to be accepted through January 2011.

The institute is committed to increasing the representation of minorities, women and individuals with physical disabilities in Computer Science. We particularly encourage such individuals to apply.

Mississippi State University

Faculty Position in Computer Science or *Software Engineering*

The Department of Computer Science and Engineering (http://www.cse.msstate.edu) is seeking to fill an open position for a tenure-track faculty member at the Assistant/Associate Professor levels. Evidence of strong potential for excellence in research (including the ability to attract external funding) and teaching at the graduate and undergraduate levels is required. The primary areas of interest for this position are Software Engineering, Artificial Intelligence, and Bioinformatics.

Mississippi State University has approximately 1300 faculty and 20,000 students. The Department of Computer Science and Engineering has 16 tenure-track faculty positions and offers academic programs leading to the bachelor's, master's and doctoral degrees in computer science and bachelor's degrees in software engineering and computer engineering. Faculty members and graduate students work with a number of on-campus research centers including the Critical Infrastructure Protection Center, the High Performance Computing Collaboratory, the Institute for Neurocognitive Science and Technology, the Institute for Digital Biology, the Center for Advanced Vehicular Systems, and the GeoResources Institute. Department research expenditures total around 5.2 million dollars per year.

Candidates for this position are expected to hold a PhD in computer science or closely related field (ABDs may be considered). Level of appointment is commensurate with qualifications and experience.

Applicants should submit a letter of application, curriculum vita, teaching statement, research statement, and names and contact information of at least three references online at http://www.jobs.msstate.edu/. Review of applications will begin not earlier than December 2010 and continue until the position is filled. MSU is an Affirmative Action/Equal Opportunity Employer.

Montana State University

RightNow Technologies Professorships in Computer Science

The Montana State University Computer Science Department is searching for two faculty members at either the Assistant, Associate or Full level, based on experience. Candidates at the Associate or Full level must have established or rising prominence in their field. A three-year start-up package is being provided by RightNow Technologies. Montana State University is a Carnegie Foundation RU/VH research university with an enrollment of approximately 13,000. The website www.cs.montana.edu/faculty-vacancies has information on position requirements and application procedures. ADA/EO/AA/Veterans Preference.

Northeastern Illinois University
Tenure-track Position

The Computer Science Department of Northeastern Illinois University in Chicago invites applications for a tenure-track position starting Fall 2011. Ph.D. in Computer Science or closely related field required. Preference will be given to candidates with interests in informatics, software engineering, emerging technologies, and computer networks and security. Strong candidates in other areas will be considered. AA/EOE View complete job posting at: http://bit.ly/neiuFac10

Oregon State University
School of Electrical Engineering and Computer Science
Two tenure-track Professorial positions in Computer Science

The School of Electrical Engineering and Computer Science at Oregon State University invites applications for two tenure-track professorial positions in Computer Science. Exceptionally strong candidates in all areas of Computer Science are encouraged to apply. We are building research and teaching strengths in the areas of open source software, internet and social computing, and cyber security, so our primary need is for candidates specializing in software engineering, database systems, web/distributed systems, programming languages, and HCI. Applicants should demonstrate a strong commitment to collaboration with other research groups in the School of EECS, with other departments at Oregon State University, and with other universities.

The School of EECS supports a culture of energetic collaboration and faculty are committed to quality in both education and research. With 40 tenure/tenure-track faculty, we enroll 160 PhD, 120 MS and 1200 undergraduate students. OSU is the only Oregon institution recognized for its "very high research activity" (RU/VH) by the Carnegie Foundation for the Advancement of Teaching. The School of EECS is housed in the Kelley Engineering Center, a green building designed to support collaboration among faculty and students across campus. Oregon State University is located in Corvallis, a college town renowned for its high quality of life.

For more information, including full position announcement and instructions for application, visit: http://eecs.oregonstate.edu/faculty/openings.php.

OSU is an AAEOE.

Peking University
School of EECS
Tenure-track Faculty Positions

The School of EECS at Peking University invites applications for tenure-track positions in the areas of energy efficient computing (including but not limited to energy-efficient computing and communication architectures, compilation, and system software) and applications (such as smart grid, mobile computing, sensor networks, and hardware acceleration of computing-intensive applications). These positions are associated with the newly established Center for Energy-Efficient Computing and Applications (http://ceca.pku.edu.cn), which offers a new level of startup and compensation packages. Applications from distinguished candidates at senior levels are also encouraged.

To apply, please email the resume, statements of research and teaching, and at least three names for references to Dr. Tao Wang wangtao@pku.edu.cn or wangtao@ieee.org. Applications received by January 15, 2011 will be given full consideration.

Princeton University
Computer Science
Assistant Professor
Tenure-Track Positions

The Department of Computer Science at Princeton University invites applications for faculty positions at the Assistant Professor level. We are accepting applications in all areas of Computer Science.

Applicants must demonstrate superior research and scholarship potential as well as teaching ability. A PhD in Computer Science or a related area is required.

Successful candidates are expected to pursue an active research program and to contribute significantly to the teaching programs of the department. Applicants should include a resume contact information for at least three people who can comment on the applicant's professional qualifications.

There is no deadline, but review of applications will start in December 2010; the review of applicants in the field of theoretical computer science will begin as early as October 2010.

Princeton University is an equal opportunity employer and complies with applicable EEO and affirmative action regulations You may apply online at:

http://www.cs.princeton.edu/jobs
Requisition Number: 1000520

Purdue University
School of ECE, Computer Engineering
Faculty Position in Human-Centered Computing

The School of Electrical and Computer Engineering at Purdue University invites applications for a faculty position at any level in human-centered computing, including but not limited to visualization, visual analytics, human computer interaction (HCI), and graphics. The Computer Engineering Area of the school (http://engineering.purdue.edu/ECE/Research/Areas/CompEng) has nineteen faculty members who have active research programs in areas including AI, architecture, compilers, computer vision, distributed systems, embedded systems, graphics, haptics, HCI, machine learning, multimedia systems, networking, networking applications, NLP, OS, robotics, software engineering, and visualization. Eligible candidates are required to have a PhD in computer science/engineering or a related field and a significant demonstrated research record commensurate with the level of the position applied for. Academic duties of the position include teaching, advising students, and maintaining a strong research program. Applications should consist of a cover letter, a CV, a research statement, names and contact information for at least three references, and URLs for three to five online papers.

Applications should be submitted to: https://engineering.purdue.edu/Engr/AboutUs/Employment/Applications.

Review of applications will begin on 1 December 2010. Inquiries may be sent to

ece-hcc-search@ecn.purdue.edu.

Applications will be considered as they are received, but for full consideration should arrive by 1 January 2011.

Purdue University is an equal opportunity, equal access, affirmative action employer fully committed to achieving a diverse workforce.

Shell
Principal Researcher - U22232

For the general field of Computational Modeling and Computer Science - To seek out, evaluate, discover and invent emerging technologies that can solve targeted current or future problems in Royal Dutch Shell (RDS) businesses.

The position is designed principally to ensure that RDS maximizes its use and leverage of external inventiveness and development, through early recognition of potentially valuable new technologies, to pursue the prospective ones, and to funnel them into the organization(s) in RDS that are the most likely to derive value.

To test ideas through the "fastest route to failure" and/or to champion and develop such technologies to the point of organizational commitment. Typically this will be to the point of investment in a funded project within the Innovation Research and Development (P&T IRD) Department, but may also be direct sponsorship by the business or external development with third parties.

The incumbent will have and maintain a broad and deep technical network with research institutes, academia, and peers in RDS and other companies including in order to be able to early recognize new ideas. The incumbent will also provide technical perspective for management on mid and long term developments in technology in the field and their impact on the valuation of business opportunities as input to technology and business strategy.

The incumbent will be expected to be a leader with credibility in the technical community and time will be made available for such a role (e.g. as active researcher and/or as PTE/SME) for project work, internal consulting or impact external representation.

Position Accountabilities:
Enhance Shell's reputation in technology and innovation through selected external representation and via purposeful interaction with external institutions universities and companies.

Leadership with respect to developing real technology options for medium and long-term technology opportunities in targeted technology focus areas such as computer science, high-performance computing, computational chemistry and physics, stochastics & computational uncertainty management, computational fluid dynamics, reservoir modeling and simulation.

Provide the future technology view for Shell Leadership as input to Strategic Decisions.

Assist with maintenance and management of a portfolio of internal and external projects on topics in fundamental research and technology that are outside the realm of existing Shell businesses or technology projects in order to rapidly and cost effectively assess their potential impact on Shell.

Accountable for the development and execution of cross business innovative R&D programs on behalf of and sponsored by Upstream and Downstream Research.

Pro-actively build/maintain a national and international network with universities, private and government research institutes, and industrial companies so that RDS does not miss the next technology wave in relevant areas.

Requirements:

▸ Legal authorization to work in the US on a full time basis for anyone other than current employer
▸ Master's in Computer Science or Computer Engineering
▸ PhD in Computer Science or Computer Engineering preferred
▸ Minimum five years (post graduate) hands on technical and project lead experience in the area of simulation, computation and modeling
▸ Minimum ten (10) professional peer reviewed or conference papers and/or patents
▸ Experience teaching and training students and/or technical professionals
▸ Proven experience running multi location/institution research projects
▸ Evidence of building and maintaining relations with leading academic and research institutes
▸ Demonstrated experience collaborating with Government agencies and/or National laboratories
▸

Application Deadline: Saturday 06 November 2010
Apply Online at
http://impact-gs.jobstreet.com/jobs/jobdesc.aspx?eid=Yf68ImbLavWpP3ctWsft6AkAkqU%3d&uid=469%7c22232%7c%7c&did=0&its=0&src=8&ref=&cc=US&agn=

Siena College
Tenure-track Faculty

The Computer Science Department at Siena College invites applicants for two tenure track positions beginning September 2011. The first position requires a Ph.D. in Computer Science with an interest in teaching a broad range of undergraduate courses. The second position requires a Ph.D. in either Computer Science or Information Systems with a very strong interest in teaching management information systems and a strong programming background. Visit http://www.siena.edu/cssearch for details on how to apply and information about the position. Salary and benefits are competitive. Siena College is an EOE.

Swarthmore College
Visiting Assistant Professor

Swarthmore College invites applications for a three-year faculty position in Computer Science, at the rank of Visiting Assistant Professor, beginning September 2011. Specialization is open. Review of applications will begin January 1, 2011, and continue until the position is filled. For information, see http://www.cs.swarthmore.edu/job.

Swarthmore College has a strong commitment to excellence through diversity in education and employment and welcomes applications from candidates with exceptional qualifications, particularly those with demonstrable commitments to a more inclusive society and world.

Texas State University-San Marcos
Department of Computer Science

Applications are invited for a tenure-track position at the rank of Assistant Professor, Associate Professor, or Professor. Consult the department recruiting page at http://www.cs.txstate.edu/recruitment/ for job duties, qualifications, application procedures, and information about the university and the department.

Texas State University-San Marcos will not discriminate against any person (or exclude any person from participating in or receiving the benefits of any of its activities or programs) on any basis prohibited by law, including race, color, age, national origin, religion, sex or disability, or on the basis of sexual orientation. Texas State University-San Marcos is a member of the Texas State University System.

The Citadel
Department of Mathematics and Computer
Assistant Professor

The Department of Mathematics and Computer Science invites applications for a tenure-track faculty position in computer science at the Assistant Professor level beginning August 2011. Qualifications include an earned Ph.D. in computer science or a closely related field and a strong commitment to excellence in teaching, research, and service. Candidates from all areas of computer science are encouraged to apply. For application procedures and information about the school and department see http://www.mathcs.citadel.edu/cs_position_2011.html.

Review of applications will begin on January 10, 2011, and will continue until the position is filled.

Applications from women and minorities are especially encouraged. The Citadel is an affirmative action/equal opportunity employer actively committed to ensuring diversity in all campus employment.

The College of William & Mary
Faculty Position in Computer Science

We invite applications for a tenure-track assistant professor position in Computer Science for Fall 2011. We are interested in individuals with research expertise in software engineering, programming languages, or compiler construction, but exceptional applicants from other areas of computer science will be considered.

The College of William & Mary, consistently ranked in the elite group of the Best National Universities-Doctoral by U.S. News and World Report, has committed to a multi-year effort to strengthen its computer science research program and, as a consequence, the department has been the home of multiple NSF CAREER Awards. Teaching loads are competitive with top research computer science departments in support of the department's research expectations for faculty. The department currently consists of thirteen faculty members who support B.S., M.S., and Ph.D. programs. More information about the department and university can be obtained at http://www.cs.wm.edu

Applicants should submit a current resume, research and teaching statements, and the names of at least three references. (Post your reference list under "other doc") Applicants must apply using William & Mary's online recruitment system at: https://jobs.wm.edu.

Review of applications will begin December 15 and continue until the position is filled. The College is an EEO/AA employer.

Applicants must hold a Ph.D. in computer science or a related field at the time of appointment, must have a strong research record, and should have an interest in teaching.

The George Washington University
Department of Computer Science
Four Faculty Positions

The Department of Computer Science at The George Washington University is seeking applicants for four faculty positions in the areas of security, systems, and AI. The first is a tenured senior position in security, at the rank of Associate or Full Professor. The other three are tenure-track positions at the rank of Assistant or Associate Professor, one in security and the other two in systems and Artificial Intelligence. People with applied or theoretical research interests in those areas are encouraged to apply. Successful candidates may start as early as Fall 2011.

Basic Qualifications: All applicants must have a Ph.D. degree in Computer Science or a closely related field. Applicants for the Associate Professor rank must have well-established and well funded research programs, and applicants for the tenured senior position in security must also be recognized scholars prepared to take on a leading research role within the department and in the field. Applicants for the Assistant Professor rank must demonstrate potential for developing a quality research program and for attracting research funding. ABD candidates for the junior positions may apply for the Assistant Professor rank, but they must complete their Ph.D. degree by August 15, 2011. All applicants must have demonstrated teaching excellence or potential at both undergraduate and graduate levels.

The George Washington University is the largest academic institution in the nation's capital with close access to many Federal funding agencies and research laboratories. The University offers comprehensive programs of undergraduate and graduate liberal arts studies as well as degrees in engineering, law, medicine, public health, education, business and international affairs. A private institution, GW prides itself on excellent research, quality education, and low student-teacher ratio. The exceptional location affords the GW community unique cultural and intellectual opportunities. In the high-tech sector, the Washington, DC Metropolitan area is one of the largest IT areas in the nation, putting us in the center of activities such as security and biotechnology.

The Department of Computer Science offers an accredited Bachelor of Science program, a Bachelor of Arts program, and Master's and Ph.D degrees. The Department has 18 faculty members, numerous affiliated and adjunct faculty members, and over 425 students. The Department has educational and research programs in security, systems, networks, graphics, biomedical applications, AI, search, and human computer interaction, with funding from various agencies; a

NSA-designated Center of Academic Excellence and Center of Academic Excellence-Research in security, with funding from NSF, DOD, and other agencies; and NIH-funded collaborations with the medical school in the biomedical areas. For further information please refer to http://www.cs.gwu.edu.

Application Procedure: To be considered, applicants must email an application to cssearch@gwu.edu containing (i) a brief cover letter that clearly indicates the position area and rank of interest, (ii) a curriculum vita, (iii) a statement of research and teaching interests, (iv) complete contact information for at least three references, and optionally (v) other relevant supporting materials. If by regular mail, applications should be sent to: Chair, Faculty Search Committee, Department of Computer Science, PHIL 703, The George Washington University, Washington D.C. 20052. Only complete applications will be considered. Inquiries about applying will be accorded the utmost discretion. Review of applications will begin on December 4, 2010, and will continue through the Spring 2011 semester, until the position is filled. For complete instructions on the application process, please visit the department faculty search website at http://www.cs.gwu.edu/facsearch/.

The George Washington University is an equal opportunity/affirmative action employer.

The IMDEA Software Institute

Tenured and Tenure-Track Faculty Positions
Madrid Institute for Advanced Studies in
Software Development Technologies (The IMDEA Software Institute)

Open call for Tenured and Tenure-track Research positions

The Madrid Institute for Advanced Studies in Software Development Technologies (The IMDEA Software Institute) invites applications for tenure-track (Research Assistant Professor) and tenured (Research Associate Professor and Research Professor) faculty positions. We are primarily interested in recruiting excellent candidates in the following areas.

- Experimental software systems, operating systems, compilers, and runtime systems.
- Rigorous empirical software engineering, including rigorous approaches to validation and testing.
- Cyber-physical systems, embedded systems, and reactive systems.
- Multicore systems and distributed computing, including cloud-computing and service-oriented architectures.
- Design and analysis of algorithms for social networks and electronic markets.

Excellent candidates in areas of established strengths of the institute, such as programming languages, program analysis, security, and verification are also encouraged to apply.

The primary mission of The IMDEA Software Institute is to perform research of excellence at the highest international level in the area of software development technologies and, in particular, to develop tools and techniques which will allow the cost-effective development of sophisticated software products with high quality, i.e., which are safe, reliable, and efficient.

Selection Process

The main selection criteria will be the candidate's demonstrated ability and commitment to research, the match of interests with the institute's mission, and how the candidate complements areas of established strengths of the institute. All positions require an earned doctoral degree in Computer Science or a closely related area. Candidates for tenure-track positions will have shown exceptional promise in research and will have displayed an ability to work independently as well as collaboratively. Candidates for tenured positions must possess an outstanding research record, have recognized international stature, and demonstrated leadership abilities.

Application materials are available at the URL
https://www.imdea.org/internationalcall/Default.aspx?IdInstitute=17

For full consideration, complete applications must be received by December 15, 2010 (and applicants should arrange for the reference letters to arrive by that date), although applications will continue to be accepted until the positions are filled.

Salaries

Salaries at The IMDEA Software Institute are internationally competitive and are established on an individual basis within a range that guarantees fair and attractive conditions with adequate and equitable social security provision in accordance with existing national Spanish legislation. This includes access to an excellent public healthcare system.

Work Environment

The working language at the institute is English. The institute is located in the vibrant area of Madrid, Spain. It offers an ideal working environment, open and collaborative, where researchers can focus on developing new ideas and projects. A generous startup package is offered. Researchers are also encouraged to participate in national and international research projects.

For more information please visit the web pages of The IMDEA Software Institute at www.software.imdea.org

IMDEA is an Equal Opportunity Employer and strongly encourages applications from a diverse and international community. IMDEA complies with the European Charter for Researchers.

The Ohio State University

Department of Computer Science and Engineering (CSE)
Assistant Professor

The Department of Computer Science and Engineering (CSE), at The Ohio State University, anticipates significant growth in the next few years. This year, CSE invites applications for four tenure-track positions at the Assistant Professor level. Priority consideration will be given to candidates in database systems, graphics & animation, machine learning, and networking. Outstanding applicants in all CSE areas (including software engineering & programming languages, systems, and theory) will also be considered.

The department is committed to enhancing faculty diversity; women, minorities, and individuals with disabilities are especially encouraged to apply.

Applicants should hold or be completing a Ph.D. in CSE or a closely related field, have a commitment to and demonstrated record of excellence in research, and a commitment to excellence in teaching.

To apply, please submit your application via the online database. The link can be found at: http://www.cse.ohio-state.edu/department/positions.shtml

Review of applications will begin in November and will continue until the positions are filled.

The Ohio State University is an Equal Opportunity/Affirmative Action Employer.

The University of Michigan, Ann Arbor

Department of Electrical Engineering and Computer Science
Computer Science and Engineering Division
Faculty Position

Applications and nominations are solicited for a faculty position in the Computer Science and Engineering (CSE) Division as part of an interdisciplinary cluster hire funded by the University President to strengthen expertise in the area of petascale computing. We are looking for individuals broadly interested in parallel systems that scale to petascale and beyond. Relevant areas include run-time systems, compilers, algorithms, programming languages, tools, and networking.

Candidates with a focus in this area are encouraged to apply. However, all computer science and engineering applications will be considered. Applications must be received by January 10, 2011.

Qualifications include an outstanding academic record, a doctorate or equivalent in computer engineering or computer science, and a strong commitment to teaching and research.

To apply please complete the form at: http://www.eecs.umich.edu/eecs/jobs/csejobs.html

Electronic applications are strongly preferred, but you may alternatively send resume, teaching statement, research statement and names of three references to:

Professor Satinder Singh Baveja, Chair, CSE Faculty Search
Department of Electrical Engineering and Computer Science
University of Michigan
2260 Hayward Street
Ann Arbor, MI 48109-2121

The University of Michigan is a Non-Discriminatory/Affirmative Action Employer with an Active Dual-Career Assistance Program. The college is especially interested in candidates who can contribute, through their research, teaching, and/or service, to the diversity and excellence of the academic community.

The University of Michigan, Ann Arbor

Dept. of Electrical Engineering and Computer Science
Computer Science and Engineering Division
Faculty Position

The University of Michigan is conducting a search for an interdisciplinary "cluster" of five new faculty over the next two or three years in the broad area of Computational Media and Interactive Systems. We

are looking for individuals whose work focuses on the interplay between computational technologies and the creative disciplines, on the use of social media and computation in arts and performance and craftsmanship, as well as on the development of new computational techniques for understanding, shaping, and building with digital media. These new positions are collectively supported by the College of Engineering (2 positions in Computer Science and Engineering), the School of Music, Theatre & Dance (1 position in Performing Arts Technology), the School of Art & Design (1 position), and Taubman College of Architecture and Urban Planning (1 position). Independently based in different disciplines and academic units, these new faculty--creative practitioners and computational engineers and scientists--will collaboratively engage innovative research and pedagogical development to explore new dimensions of creativity made possible by a shared digital ecology.

As part of the interdisciplinary cluster hiring, applications are solicited for a faculty position in the Computer Science and Engineering (CSE) Division in all areas related to analyzing, understanding, representing, and creating computational media, including but not limited to audio, music, image, and video information retrieval, video understanding and mashups, time-based, distributed, collaborative and interactive media, and automatic classification, clustering, and activity recognition in heterogeneous media.

Qualifications include an outstanding academic record, a doctorate or equivalent in computer science or computer engineering or a discipline relevant to the position, and a strong commitment to teaching and research.

To apply please complete the form at: http://www.eecs.umich.edu/eecs/jobs/csejobs.html

Electronic applications are strongly preferred, but you may alternatively send resume, teaching statement, research statement and names of three references to:

Professor Satinder Singh Baveja
Chair, CSE Faculty Search
Department of Electrical Engineering and Computer Science
University of Michigan
2260 Hayward Street
Ann Arbor, MI 48109-2121

The University of Michigan is a Non-Discriminatory/Affirmative Action Employer with an Active Dual-Career Assistance Program. The college is especially interested in candidates who can contribute, through their research, teaching, and/or service, to the diversity and excellence of the academic community.

Towson University

Assistant Professor

The Department of Computer and Information Sciences at Towson University invites applications for two tenure-track positions in its Computer Science (CS) and Information Technology (IT) programs. Selected candidates will be expected to teach undergraduate and graduate courses, participate in department activities, conduct research, and supervise graduate students. Applicants for the CS position must hold a Ph.D. in Computer Science, have a strong publication record, and the potential of attracting external funding; all areas of research will be considered. Applicants for the IT position must hold a Ph.D. in an IT-related field, with those having industry experience and the potential for applied research funding encouraged to apply.

Applicants should electronically submit a letter of application, current resume, copies of graduate transcripts, a recent research paper, and the name, address, phone and e-mail address of three professional references to **CSSearch@towson.edu** (for the CS position) or **ITSearch@towson.edu** (for the IT position). Applicants should only apply for one of the two positions.

Full position announcements available at:
http://wwwnew.towson.edu/odeo/employmentatTU/AssistantProfessorComputerScience.asp
http://wwwnew.towson.edu/odeo/employmentatTU/AssistantProfessorInformationTechnology.asp

The review of completed applications will begin January 15, 2011 and continue until the positions are filled.

Towson University is an equal opportunity/affirmative action employer and has a strong institutional commitment to diversity. Women, minorities, persons with disabilities and veterans are encouraged to apply.

The University of Michigan - Dearborn

Department of Computer and Information Science

Tenure-Track Positions

The Department of Computer and Information Science (CIS) at the University of Michigan-Dearborn invites applications for two tenure-track faculty positions. Rank and salary will be commensurate with qualifications and experience. We offer competitive salaries and start-up packages.

Position #1

Applications for an Assistant/Associate/Full Professor position in any of the following areas: computer and data security, digital forensics, and information assurance.

Position #2

Applications for an Assistant/Associate Professor position. While all areas will be considered, we will put special emphasis on applicants in the following areas: distributed information-centric systems (sensor networks/databases, distributed information systems, data dissemination, cloud computing, network management and social networks), semantic computing (semantic web, web services, ontology engineering, semantic-enabled applications, semantic integration, semantic interfaces, semantics-based analysis).

Qualified candidates must have, or expect to have, a Ph.D. in CS or a closely related discipline by the time of appointment and will be expected to do scholarly and sponsored research, as well as teaching at both the undergraduate and graduate levels. Candidates at the associate or full professor ranks should already have an established funded research program. The CIS Department, currently at 14 tenure-track faculty, offers several BS and MS degrees, and participates in an interdisciplinary Ph.D. program in information systems engineering. The current research areas in the department include computer graphics and geometric modeling, database and information management, multimedia systems and gaming, networking, computer and network security, and software engineering. These areas of research are supported by several established labs and sponsored by many external funding agencies.

The University of Michigan-Dearborn is located in the southeastern Michigan area and offers excellent opportunities for faculty collaboration with many industries. We are one of three campuses forming the University of Michigan system and are a comprehensive university with over 8500 students. One of university's strategic visions is to advance the future of manufacturing in a global environment.

The University of Michigan-Dearborn is dedicated to the goal of building a culturally-diverse and pluralistic faculty committed to teaching and working in a multicultural environment, and strongly encourages applications from minorities and women.

A cover letter including the position number under consideration, curriculum vitae including e-mail address, teaching statement, research statement, and three letters of reference should be sent to:

Dr. William Grosky, Chair
Department of Computer and Information Science
University of Michigan-Dearborn
4901 Evergreen Road
Dearborn, MI 48128-1491
Email: wgrosky@umich.edu,
Internet: http://www.cis.umd.umich.edu
Phone: 313.583.6424, Fax: 313.593.4256

The University of Michigan-Dearborn is an equal opportunity/affirmative action employer.

Toyota Technological Institute at Chicago

Computer Science Faculty Positions at All Levels

Toyota Technological Institute at Chicago (TTIC) is a philanthropically endowed degree-granting institute for computer science located on the University of Chicago campus. The Institute is expected to reach a steady-state of 12 traditional faculty (tenure and tenure track), and 12 limited term faculty. Applications are being accepted in all areas, but we are particularly interested in

Theoretical computer science
Speech processing
Machine learning
Computational linguistics
Computer vision
Computational biology
Scientific computing

Positions are available at all ranks, and we have a large number of limited term positions currently available.

For all positions we require a Ph.D. Degree or Ph.D. candidacy, with the degree conferred prior to date of hire. Submit your application electronically at:

http://ttic.uchicago.edu/facapp/

Toyota Technological Institute at Chicago is an Equal Opportunity Employer.

United States Naval Academy
Computer Science Department
Assistant Professor

The U.S. Naval Academy's Computer Science Department invites applications for one or more tenure track positions at the rank of Assistant Professor. These positions are anticipated to begin in the Autumn of 2011. A Ph.D. in Computer Science or closely related field is required.

The Computer Science Department offers ABET accredited majors in Computer Science and Information Technology. All faculty teach courses in both majors. We currently have 80 CS majors, 90 IT majors and a faculty of 15. In the summer of 2004, the department moved into a newly renovated building overlooking the scenic Severn River. Our space provides outstanding office, laboratory, and research facilities for both students and faculty, including specialized labs for robotics, networking, and information assurance in addition to three micro-computing labs and two high performance computing labs.

The Naval Academy is an undergraduate institution located in historic downtown Annapolis, MD on the Chesapeake Bay. Roughly half of its faculty are tenured or tenure track civilian professors with Ph.D.s who balance teaching excellence with internationally recognized research programs. The remaining faculty are active duty military officers with Masters or Doctoral degrees. Each year the academy graduates roughly 1000 undergraduate students with majors in the sciences, engineering, and humanities. More information about the department and the Academy can be found at http://www.usna.edu/cs/ and http://www.usna.edu/.

Applicants must have a dedication to teaching, broad teaching interests, and a strong research program. Applications will be considered from all areas of Computer Science.

Applicants should send a cover letter, teaching and research statements, a curriculum vitae, and three letters of recommendation that address both teaching and research abilities to cssearch@usna.edu.

Review of applications will begin November 1, continuing until the position is filled.

The United States Naval Academy is an Affirmative Action/Equal Opportunity Employer. This agency provides reasonable accommodations to applicants with disabilities. This position is subject to the availability of funds.

University of California, Los Angeles
Computer Science Department

The Computer Science Department of the Henry Samueli School of Engineering and Applied Science at the University of California, Los Angeles, invites applications for tenure-track positions in all areas of Computer Science and Computer Engineering. Applications are also encouraged from distinguished candidates at senior levels. Quality is our key criterion for applicant selection. Applicants should have a strong commitment both to research and teaching and an outstanding record of research for their level of seniority. Salary is commensurate with education and experience.

UCLA is an Equal Opportunity/Affirmative Action Employer. The department is committed to building a more diverse faculty, staff and student body as it responds to the changing population and educational needs of California and the nation. To apply, please visit http://www.cs.ucla.edu/recruit. Faculty applications received by January 15 will be given full consideration.

University of Calgary
Department of Computer Science
Assistant Professor Positions

The Department of Computer Science and the University of Calgary seeks outstanding candidates for two tenure track positions at the Assistant Professor level. Applicants from areas of Database Management and Human Computer Interaction/Information Visualization are of particular interest. Details for each position appear at: http://www.cpsc.ucalgary.ca/.

Applicants must possess a doctorate in Computer Science or a related discipline at the time of appointment, and have a strong potential to develop an excellent research record.

The department is one of Canada's leaders as evidenced by our commitment to excellence in research and teaching. It has an expansive graduate program and extensive state-of-the-art computing facilities. Calgary is a multicultural city that is the fastest growing city in Canada. Calgary enjoys a moderate climate located beside the natural beauty of the Rocky Mountains. Further information about the department is available at http://www.cpsc.ucalgary.ca/.

Interested applicants should send a CV, a concise description of their research area and program, a statement of teaching philosophy, and arrange to have at least three reference letters sent to:

Dr. Carey Williamson
Department of Computer Science
University of Calgary
Calgary, Alberta, Canada T2N 1N4 or
search@cpsc.ucalgary.ca

The applications will be reviewed beginning November 2010 and continue until the positions are filled.

All qualified candidates are encouraged to apply; however, Canadians and permanent residents will be given priority. The University of Calgary respects, appreciates, and encourages diversity.

University of Delaware
Department of Computer and Information Sciences
Network/Systems Security Faculty Position

Applications are invited for a tenure-track faculty position at all ranks in Network/Systems Security to begin Fall 2011.

More information is available at https://www.engr.udel.edu/faculty-search.

University of Delaware
Faculty Positions in National Security Technologies

The College of Engineering at the University of Delaware invites nominations and applications for all levels with mid- and senior-level tenure-track faculty positions particularly encouraged to lead a growing program in National Security Technology (www.engr.udel.edu/forms/facultysearch/); exceptional junior-level applications will also be considered. Candidates with backgrounds in all engineering fields, with research interests in technologies related to the protection and preservation of the US national interest both within and outside our boarders are particularly encouraged to apply.

Appointments may be in a primary engineering discipline or as an interdisciplinary appointment across departments. Candidates will be expected to teach undergraduate and graduate classes within their discipline and to conduct innovative and internationally recognized research. If appropriate, candidates applying in areas requiring a high-level security clearance should either possess or be capable of attaining one within the Department of Defense or equivalent Federal Agency.

Applicants should submit a curriculum vitae, a statement of research and teaching interests and achievements, and the names, addresses, phone numbers, and e-mail addresses of four references at www.engr.udel.edu/forms/facultysearch/. Review of applications will begin as early as October 15, 2010, although nominations and applications will be accepted until the position is filled.

The UNIVERSITY OF DELAWARE is an Equal Opportunity Employer which encourages applications from Minority Group Members and Women.

University of Delaware
Faculty Positions in Computer and Information Technology

The College of Engineering at the University of Delaware invites nominations and applications for all levels with mid- and senior-level tenure-track faculty positions particularly encouraged to lead a growing program in National Security Technology (www.engr.udel.edu/forms/facultysearch/); exceptional junior-level applications will also be considered. Candidates with backgrounds in all engineering fields, with research interests in technologies related to the protection and preservation of the US national interest both within and outside our boarders are particularly encouraged to apply.

Appointments may be in a primary engineering discipline or as an interdisciplinary appointment across departments. Candidates will be expected to teach undergraduate and graduate classes within their discipline and to conduct innovative and internationally recognized research. If appropriate, candidates applying in areas requiring a high-level security clearance should either possess or be capable of attaining one within the Department of Defense or equivalent Federal Agency.

Applicants should submit a curriculum vitae, a statement of research and teaching interests and achievements, and the names, addresses, phone numbers, and e-mail addresses of four references at www.engr.udel.edu/forms/facultysearch/. Review of applications will begin as early as October 15, 2010, although nominations and applications will be accepted until the position is filled.

The UNIVERSITY OF DELAWARE is an Equal Opportunity Employer which encourages applications from Minority Group Members and Women.

University of Illinois Springfield
Assistant Professor of Computer Science

The Computer Science Department at the University of Illinois at Springfield (UIS) invites applications for a beginning assistant professor, tenure track position to begin August, 2011. Please note that a Ph.D. in Computer Science or closely related field is required at the time of hire. The position involves graduate and undergraduate teaching, supervising student research, and continuing your research. Many of our classes are taught online. All areas of expertise will be considered, but computer security is of special interest to the Department. Review of applications will begin December 1, 2010 and continue until the position is filled or the search is terminated. Please send your vita and contact information for three references to Chair Computer Science Search Committee; One University Plaza; UHB 3100; Springfield, IL 62703-5407.

Located in the state capital, the University of Illinois Springfield is one of three campuses of the University of Illinois. The UIS campus serves approximately 5,100 students in 20 graduate and 23 undergraduate programs. The academic curriculum of the campus emphasizes a strong liberal arts core, an array of professional programs, extensive opportunities in experiential education, and a broad engagement in public affairs issues of the day. The campus offers many small classes, substantial student-faculty interaction, and a rapidly evolving technology enhanced learning environment. Its diverse student body includes traditional, non-traditional, and international students. Twenty-five percent of majors are in 17 undergraduate and graduate online degree programs and the campus has received several national awards for its implementation of online learning. UIS faculty are committed teachers, active scholars, and professionals in service to society. You are encouraged to visit the university web page at http://www.uis.edu and the department web page at http://csc.uis.edu . UIS is an affirmative action/equal opportunity employer with a strong institutional commitment to recruitment and retention of a diverse and inclusive campus community. Women, minorities, veterans, and persons with disabilities are encouraged to apply.

University of Maryland College Park
Department of Computer Science

The Department of Computer Science at the University of Maryland, College Park, MD, USA has two openings for faculty positions effective July 1, 2011 or earlier. Applicants will be considered for joint appointments between the Department and the Institute for Advanced Computer Studies (UMIACS). We invite applications from candidates to fill two positions at the junior or senior levels (tenure-track Assistant Professor or tenured Associate or Full Professor) in Cyber Security or a related area.

Applications from women and minority candidates are especially welcome.
Please apply online at https://jobs.umd.edu and hiring.cs.umd.edu. Candidates must apply to both websites to receive consideration. Applications should be completed by January 3, 2011 for best consideration. The review of applicants will be on-going, so we encourage your early application.

Additional information about the Department of Computer Science and the Institute for Advanced Computer Studies is available at http://www.cs.umd.edu and at http://www.umiacs.umd.edu .

The University of Maryland is an Equal Opportunity, Affirmative Action Employer.

University of Massachusetts Amherst
Assistant Professor, Computer Science, Tenure Track, Machine Learning

We invite applications for a tenure-track Assistant Professor in the area of machine learning, especially graphical models and the development of novel machine learning methodology for a diverse range of application areas. Applicants must have completed (or be completing) a Ph.D. in Computer Science, or a related area, and should show evidence of exceptional research promise.

Our department, which has a strong record of interdisciplinary collaboration, has prominent and growing strengths in various areas of machine learning and AI. In addition, many of our faculty work in areas neighboring machine learning, including vision, robotics, natural language, information retrieval, data mining, computational social science, sensor networks and computer networks. We are highly supportive of junior faculty, providing both formal and informal mentoring. We have a substantial history of NSF CAREER awards and other early research funding. Computer Science has close ties to other departments including statistics/mathematics, engineering, biology, physics, linguistics and the social sciences, as well as new "green" initiatives. Amherst, a historic New England town, is the center of a vibrant and culturally rich area that includes four other colleges.

To apply, please send a cover letter referencing search R39789 with your vita, a research statement, a teaching statement, and at least three letters of recommendation. Electronic submission of application materials in pdf format is preferred. Send to facrec@cs.umass.edu. Alternatively, paper copies of application materials may be sent to: Search R39789, c/o Chair of Faculty Recruiting, Department of Computer Science, University of Massachusetts, Amherst, MA 01003.

We will begin to review applications on December 1, 2010 and will continue until the position is filled. Salary will be commensurate with education and experience. Inquiries and requests for more information can be sent to: facrec@cs.umass.edu

The University of Massachusetts is an Affirmative Action/Equal Opportunity employer. Women and members of minority groups are encouraged to apply.

University of Nevada, Las Vegas - UNLV
Data Visualization Faculty Position

The Howard R. Hughes College of Engineering, University of Nevada Las Vegas (UNLV), invites applications from all engineering and computer science disciplines for a full-time tenure-track faculty position in Data Visualization at the Assistant Professor level commencing Fall 2011. A complete job description with application details may be obtained by visiting http://jobs.unlv.edu/ or call (702) 895-2894. A review of applications will begin immediately. EEO/AA Employer

University of North Florida
Director of School of Computing

The University of North Florida invites applications for the position of Director of the School of Computing. Candidates for this position must have attained the rank of Full Professor, have an earned doctorate in computing, and must be eligible for tenure within the School of Computing. Successful candidates must have administrative and managerial experience; strong record in teaching, sponsored research, and service; strong record of leadership; appreciation of interrelationships among programs; record of fostering collaboration with external constituencies; commitment to strategic planning, program assessment, and accreditation efforts.

Review of applications begins December 6, 2010 and the position is open until filled. Salary is negotiable. The anticipated start date is July 1, 2011. Applicant must complete a one-page application on-line at http://www.unfjobs.org (search position: 315250) and must submit all required documents to be considered for this position. UNF is an Equal Opportunity/Equal Access/Affirmative Action Institution.

University of Pennsylvania
Department of Computer and Information Science
Faculty Position

The University of Pennsylvania invites applicants for tenure-track appointments in computer graphics and animation to start July 1, 2011. Tenured appointments will also be considered.

Faculty duties include teaching undergraduate and graduate students and conducting high-quality research. Teaching duties will be aligned with two programs: the Bachelor of Science and Engineering in Digital Media Design, and the Master of Science and Engineering in Computer Graphics and Game Technology (see http://cg.cis.upenn.edu). Research and teaching will be enhanced by the recently renovated SIG Center for Computer Graphics, which houses the largest motion capture facility in the region, and is also the home of the Center for Human Modeling and Simulation. Successful applicants will find Penn to be a stimulating environment conducive to professional growth.

The University of Pennsylvania is an Ivy League University located near the center of Philadelphia, the 5th largest city in the US. Within walking distance of each other are its Schools of Arts and Sciences, Engineering, Fine Arts, Medicine, the Wharton School, the Annenberg School of Communication, Nursing, and Law. The University campus and the Philadelphia area support a rich diversity of scientific, educational, and cultural opportunities, major technology-driven industries such as pharmaceuticals, finance, and aerospace, as well as attractive urban and suburban residential neighborhoods. Princeton and

New York City are within commuting distance.

To apply, please complete the form located on the Faculty Recruitment Web Site at: http://www.cis.upenn.edu/departmental/facultyRecruiting.shtml. Electronic applications are strongly preferred, but hard-copy applications (including the names of at least four references) may alternatively be sent to:

Chair, Faculty Search Committee
Department of Computer and Information Science
School of Engineering and Applied Science
University of Pennsylvania
Philadelphia, PA 19104-6389.

Applications should be received by January 15, 2011 to be assured full consideration. Applications will be accepted until the position is filled. Questions can be addressed to faculty-search@cis.upenn.edu. The University of Pennsylvania values diversity and seeks talented students, faculty and staff from diverse backgrounds.

The University of Pennsylvania does not discriminate on the basis of race, sex, sexual orientation, gender identity, religion, color, national or ethnic origin, age, disability, or status as a Vietnam Era Veteran or disabled veteran in the administration of educational policies, programs or activities; admissions policies; scholarship and loan awards; athletic, or other University administered programs or employment. The Penn CIS Faculty is sensitive to "two–body problems" and opportunities in the Philadelphia region.

University of Rochester

Assistant to Full Professor of Computer Science

The UR Department of Computer Science seeks researchers in computer vision and/or machine learning for a tenure-track faculty position beginning in Fall 2011. Outstanding applicants in other areas may be considered. Candidates must have a PhD in computer science or related discipline. Senior candidates should have an extraordinary record of scholarship, leadership, and funding.

The Department of Computer Science is a select research-oriented department, with an unusually collaborative culture and strong ties to cognitive science, linguistics, and electrical and computer engineering. Over the past decade, a third of its PhD graduates have won tenure-track faculty positions, and its alumni include leaders at major research laboratories such as Google, Microsoft, and IBM.

The University of Rochester is a private, Tier I research institution located in western New York State. The University of Rochester consistently ranks among the top 30 institutions, both public and private, in federal funding for research and development. Half of its undergraduates go on to post-graduate or professional education. The university includes the Eastman School of Music, a premiere music conservatory, and the University of Rochester Medical Center, a major medical school, research center, and hospital system. The Rochester area features a wealth of cultural and recreational opportunities, excellent public and private schools, and a low cost of living.

Candidates should apply online at http://www.cs.rochester.edu/recruit after Nov. 1, 2010. Review of applications will begin on Dec. 1, and continue until all interview openings are filled. The University of Rochester has a strong commitment to diversity and actively encourages applications from candidates from groups underrepresented in higher education. The University is an Equal Opportunity Employer.

University of South Carolina

Arnold School of Public Health
Department of Exercise Science
Endowed Chair Position

The University of South Carolina (USC) seeks to hire an endowed chair of Technology Applications for Health Behavior Change. We are seeking applicants with an international reputation for delivering health behavior change interventions via innovative mass communication and cost-effective approaches. The successful applicant must have published extensively in the peer-reviewed literature in this area and must have a vision for leadership for the Center. The chair will be faculty in the Arnold School of Public Health (**http://sph.sc.edu**), Department of Exercise Science.

This endowed chair is one of two in the Technology Center to Enhance Healthful Lifestyles, a Center of Economic Excellence (CoEE). The other endowed chair position has been established at the Medical University of South Carolina. The CoEE program (**www.sccoee.org**/) is designed to generate a critical mass of researchers, businesses, and service providers in identified research fields.

The successful applicant will join a strong team of USC investigators in the areas of physical activity and dietary change, and will collaborate with the major health care systems in SC, the Department of Health and Environmental Control (DHEC), and private medical clinics and individual practitioners. He or she will be expected to obtain external funds to support innovative new interventions and to provide significant and effective mentorship of doctoral students, postdoctoral fellows, and junior faculty, and contribute to the teaching mission of the University.

Candidates are encouraged to submit a letter articulating their research leadership experience and goals, a curriculum vitae, and a statement of research interests and accomplishments to:

Professor Sara Wilcox, Chair of the Technology Applications for Health Behavior Change CoEE Search Committee, Arnold School of Public Health, 921 Assembly Street, 2nd Floor, University of South Carolina, Columbia, SC, 29208. E-Mail applications are encouraged and can be sent to **wilcoxs@mailbox.sc.edu**. A list of references will be sought after an initial review of applications.

The University of South Carolina is an affirmative action, equal opportunity employer.

Women and minorities are strongly encouraged to apply.

University of Washington

Computer Science & Engineering
Tenure-Track, Research, and Teaching Faculty

The University of Washington's Department of Computer Science & Engineering has one or more open positions in a wide variety of technical areas in both Computer Science and Computer Engineering, and at all professional levels. A moderate teaching load allows time for quality research and close involvement with students. Our space in the Paul G. Allen Center for Computer Science & Engineering provides opportunities for new projects and initiatives. The Seattle area is particularly attractive given the presence of significant industrial research laboratories as well as a vibrant technology- driven entrepreneurial community that further enhances the intellectual atmosphere. Information about the department can be found on the web at http://www.cs.washington.edu.

We welcome applicants in all research areas in Computer Science and Computer Engineering including both core and inter-disciplinary areas. We expect candidates to have a strong commitment both to research and to teaching. The department is primarily seeking individuals at the tenure-track Assistant Professor rank; however, under unusual circumstances and commensurate with the qualifications of the individual, appointments may be made at the rank of Associate Professor or Professor. We may also be seeking non-tenured research faculty at Assistant, Associate and Professor levels, postdoctoral researchers (Research Associates) and part-time and full-time annual lecturers and Sr. Lecturers. Research Associates, Lecturers and Sr. Lecturers will be hired on an annual or multi-annual appointment. All University of Washington faculty engage in teaching, research and service.

Applicants for both tenure-track and research positions must have earned a doctorate by the date of appointment; those applying for lecturer positions must have earned at least a Master's degree.

Please apply online at http://www.cs.washington.edu/news/jobs.html with a letter of application, a complete curriculum vitae, statement of research and teaching interests, and the names of four references. Applications received by December 15, 2010 will be given priority consideration. Open positions are contingent on funding.

The University of Washington was awarded an Alfred P. Sloan Award for Faculty Career Flexibility in 2006. In addition, the University of Washington is a recipient of a National Science Foundation ADVANCE Institutional Transformation Award to increase the participation of women in academic science and engineering careers. We are building a culturally diverse faculty and encourage applications from women and minority candidates. The University of Washington is an affirmative action, equal opportunity employer.

University of Waterloo

David R. Cheriton School of Computer Science
Tenured and Tenure-Track Faculty Positions

Applications are invited for several positions in computer science: (a) Up to two senior, tenured David R. Cheriton Chairs in Software Systems are open for candidates with outstanding research records in software systems (very broadly defined). Successful applicants will be acknowledged leaders in their fields or have demonstrated the potential to become such leaders. These positions include substantial research support and teaching reduction. (b) One tenured or tenure-track position is open in the area of Health Informatics, including, but not limited to, healthcare IT, medical informatics, and biomedical systems.

The successful applicant will help develop a new graduate degree program in health informatics. (c) One other tenured or tenure-track position is available for excellent candidates in any computing area, but highest priority will be given to candidates specializing in systems software (operating systems, distributed systems, networks, etc.) and information systems (e-commerce systems, enterprise resource planning systems, business intelligence, etc.).

Successful applicants who join the University of Waterloo are expected to be leaders in research, have an active graduate-student program, and contribute to the overall development of the School. A Ph.D. in Computer Science, or equivalent, is required, with evidence of excellence in teaching and research. Rank and salary will be commensurate with experience, and appointments are expected to commence during the 2011 calendar year.

With over 70 faculty members, the University of Waterloo's David R. Cheriton School of Computer Science is the largest in Canada. It enjoys an excellent reputation in pure and applied research and houses a diverse research program of international stature. Because of its recognized capabilities, the School attracts exceptionally well-qualified students at both undergraduate and graduate levels. In addition, the University has an enlightened intellectual property policy which vests rights in the inventor: this policy has encouraged the creation of many spin-off companies including iAnywhere Solutions Inc., Maplesoft Inc., Open Text Corp., and Research in Motion. Please see our web site for more information: http://www.cs.uwaterloo.ca.

To submit an application, please register at the submission site: http://www.cs.uwaterloo.ca/faculty-recruiting. Once registered, instructions will be provided regarding how to submit your application. Although applications will be considered as soon as possible after they are complete and as long as positions are available, full consideration is assured for those received by November 30.

The University of Waterloo encourages applications from all qualified individuals, including women, members of visible minorities, native peoples, and persons with disabilities. All qualified candidates are encouraged to apply; however, Canadian citizens and permanent residents will be given priority. Fall 2010.

Utah State University

Assistant Professor

Applications are invited for a faculty position at the Assistant Professor level, for employment beginning Fall 2011. Applicants must have completed a PhD in computer science by the time of appointment. The position requires demonstrated research success, a significant potential for attracting external research funding, excellence in teaching both undergraduate and graduate courses, the ability to supervise student research, and excellent communication skills. The department is interested in strengthening its focus in the following areas: Software Security, Game Development, and Database Systems.

USU offers competitive salaries and outstanding medical, retirement, and professional benefits (see http://www.usu.edu/hr/ for details). The department currently has approximately 280 undergraduate majors, 80 MS students and 27 PhD students. There are 17 full time faculty. The BS degree is ABET accredited. Utah State University is a Carnegie Research Doctoral extensive University of over 23,000 students, nestled in a mountain valley 80 miles north of Salt Lake City, Utah. Opportunities for a wide range of outdoor activities are plentiful. Housing costs are at or below national averages, and the area provides a supportive environment for families and a balanced personal and professional life. Women, minority, veteran and candidates with disabilities are encouraged to apply. USU is sensitive to the needs of dual-career couples. Utah State University is an affirmative action/equal opportunity employer, with a National Science Foundation ADVANCE Gender Equity program, committed to increasing diversity among students, faculty, and all participants in university life.

Applications must be submitted using USU's online job-opportunity system. To access this job opportunity directly and begin the application process, visit https://jobs.usu.edu/applicants/Central?quickFind=54615.

The review of the applications will begin on January 15, 2011 and continue until the position is filled. The salary will be competitive and depend on qualifications.

Valdosta State University

Head, Department of Mathematics and Computer Science

Valdosta State University is accepting applications for an administrative tenure-track, fiscal-year position as Head of the Department of Mathematics and Computer Science at the rank of associate or full professor. For a complete job vacancy announcement and application instructions email hhatcher@valdosta.edu. The starting date is July 1, 2011. The review of applications will begin on November 15, 2010, position remains open until filled. Valdosta State University is an Equal Opportunity educational institution and has a strong institutional commitment to diversity. In that spirit, we are particularly interested in receiving applications from a broad spectrum of people, including, but not limited to, minorities, and individuals with disabilities. Valdosta State University has a non-discrimination policy that includes sex, race, color, sexual orientation, religion, age, marital status, national origin, disability, and veteran status.

Requirements include a Ph.D. in mathematics, computer science, or a closely related field and strong teaching and research skills; some administrative experience preferred. Candidates must possess strong interpersonal and effective leadership skills and be committed to excellence in teaching.

Washington State University Vancouver

Tenure-Track Position, Assistant Professor level

COMPUTER SCIENCE FACULTY – Washington State University Vancouver invites applications for a tenure-track position at the assistant professor level beginning 8/16/2011. Candidates are sought with expertise in **computer networks, wireless networks or sensor networks**. Position duties include teaching, research and service.

Required qualifications: Ph.D. in Computer Science or Computer Engineering at the time of employment and demonstrated ability to (1) develop funded research program, (2) establish strong industrial collaborations, and (3) teach undergraduate/graduate courses. Preferred qualifications: knowledge of the ABET accreditation process, relevant industrial background and commitment to working with diverse student and community populations. WSU Vancouver is committed to building a culturally diverse educational environment.

WSU Vancouver serves about 3000 graduate and undergraduate students and is **fifteen miles north of Portland, Oregon**. The rapidly growing School of Engineering and Computer Science (ENCS) equally values both research and teaching. WSU is Washington's land grant university with faculty and programs on four campuses. For more information: http://encs.vancouver.wsu.edu/.

Applications must include: (1) cover letter with a clear description of experience relevant to the position; (2) vita including a list of references; and (3) **maximum three-page total** summary statement of research and teaching experience. This statement must describe how the candidate's research activity will expand or complement the current research in ENCS. It must also list the existing ENCS courses and proposed new courses the candidate can develop/teach. Application deadline is **December 15, 2010**. Submit application materials online at http://www.wsujobs.com. WSU is committed to excellence through diversity, has faculty friendly policies including a partner accommodation program, and a NSF ADVANCE Institutional Transformation grant (see http://www.excelinse.wsu.edu/). WSU employs only US citizens and lawfully authorized non-citizens. WSU is an EO/AA educator and employer.

Yale University

Senior Faculty

Yale University's Electrical Engineering Department invites applications from qualified individuals for a senior faculty position in either computer systems or signals & systems. Subfields of interest include wireless communications, networking, systems on a chip, embedded systems, and emerging computing methodologies inspired by advances in the biological sciences, quantum computing, and other novel research directions. All candidates should be strongly committed to both teaching and research and should be open to collaborative research. Candidates should have distinguished records of research accomplishments and should be willing and able to take the lead in the shaping of Yale's expanding programs in either computer engineering or signals & systems. Yale University is an Affirmative Action/Equal Opportunity Employer. Yale values diversity among its students, staff, and faculty and strongly welcomes applications from women and under represented minorities. The review process will begin November 1, 2010. Applicants should send a curriculum vitae to:

Chair
Electrical Engineering Search Committee
Yale University
P.O. Box 208284
New Haven, CT 06520-8284

DOI:10.1145/1859204.1859232

Peter Winkler

Puzzled
Solutions and Sources

It's amazing how little we know about the simple, ordinary, axis-aligned rectangle. Last month (p. 112) we posted a trio of brainteasers, including one as yet unsolved, concerning rectangles galore. Here, we offer solutions to at least two of them. How did you do?

1. Partitioning a Rectangle. *Solution.* We were given a large rectangle in the plane, partitioned into a finite number of smaller rectangles, each with either integer width or integer height. Our mission was to prove the big rectangle also has integer width or height.

The puzzle was the subject of a famous article "Fourteen Proofs of a Result About Tiling a Rectangle" by Stan Wagon of Macalester College, St. Paul, MN, in *The American Mathematical Monthly 94* (1987). Here, we provide one of several solutions not found among Wagon's 14. Letting epsilon be less than the smallest tolerance in the partition, color each small rectangle of integral width green, except for a red horizontal strip of width epsilon across the top and another across the bottom. Next, color each remaining small rectangle red, except for a green vertical strip of width epsilon along the left side and another along the right side. Now place the lower-left point of the big rectangle at the origin.

There is either a green path from the left side of the big rectangle to the right side or a red path from the bottom to the top. Suppose the former. Every time the green path crosses a vertical border of the partition, it is at an integral coordinate. The big rectangle thus has integral width. Similarly, a red path from bottom to top forces integral height.

2. Blocking the Enemy. *Solution.* Recall we were in a large rectangular room with mirrored walls, while elsewhere in the same room was our mortal enemy, armed with a laser gun. Our only defense was our ability to summon graduate students to stand at designated spots in the room blocking all possible shots by the enemy. How many students would we need? We assume for the purposes of the problem that we, our enemy, and the students are all thin enough to be considered points. So, for example, if we had continuum many students, we could place them around us in a circle (with the enemy outside). But we can do better.

This wonderful puzzle seems to have begun life at the Leningrad Mathematics Olympiad of 1990 and has since spurred some serious mathematics by, among others, Benjamin Schmidt and Jean-Francois Lafont of The Ohio State University and Eugene Gutkin of the University of Southern California.

First, view the room as a rectangle in the plane, with us at P and the enemy at Q. We can now tile the plane with copies of the room by repeatedly reflecting the room about its walls, with each copy containing a new copy of our enemy. Every possible shot by the enemy can be represented on this image by a straight line from some copy of the point Q to P. Every time such a line crosses a boundary between rectangles, the real laser beam bounces off a wall. This observation already tells us there is only a countably infinite number of ways for the enemy to shoot us, so a countable number of students is enough.

But the shots (once folded into the original room) intersect one another frequently, and, conceivably, a well-placed student could block infinitely many shots. Indeed, it takes only 16 students to block every possible shot at exactly its halfway point. To see how it works, "trace" a copy of the plane tiling onto a piece of paper, pin it to the plane at our position P, and shrink the paper copy by a factor of two vertically and horizontally. The many copies of Q on the shrunk copy will be the students' positions, serving our purpose because each copy of Q on the original tiling appears halfway between it and us on the shrunk copy. There are infinitely many such points, but all are copies of a set of 16 points in the original room.

3. Covering a Big Rectangle. *Unsolved.* The third problem is open. No one can prove we can even get the small rectangles to cover at least 1% of the big rectangle. Another mystery is its origin. I heard it 20 years ago from Bill Pulleyblank (now professor of operations research at the U.S. Military Academy, West Point, NY), though he doesn't recall where he got it.

Peter Winkler (puzzled@cacm.acm.org) is Professor of Mathematics and of Computer Science and Albert Bradley Third Century Professor in the Sciences at Dartmouth College, Hanover, NH.

All readers are encouraged to submit prospective puzzles for future columns to puzzled@cacm.acm.org.

[CONTINUED FROM P. 128] fad for string quartets and harpsichord recitals, and regularly assists the Royal Society in its scientific research on the tribes and wildlife of the region.

It was with extreme incredulity that Edmund heard there were Barbary pirates in the Caribbean; he thought they were confined to the Old World but soon learned the very concept of "Barbary pirate" was slanderous European propaganda against the North Africans, who were behaving no differently from the privateers commissioned by European governments, with their letters of marque and reprisal. *Pirates of the Burning Sea* occasionally plays with the facts of history, but always frankly, showing great respect for the truth and admiring the remarkable accomplishments of the 18th century. Yet this marvelous virtual world has few visitors.

Another educationally valuable virtual world is *A Tale in the Desert* (http://www.atitd.com), a set of nonviolent social games with an ancient Egyptian motif with perhaps only 1,000 subscribers today, one 10,000th the peak population of *World of Warcraft* (http://www.worldofwarcraft.com).

Tens of thousands of people might suddenly subscribe to both *Pirates of the Burning Sea* and *A Tale in the Desert*, though this is probably a vain hope. More realistically, leaders of government, education, and computer science in your world could establish a digital library to host the best of the early virtual worlds, not as historical curiosities, but as immortal masterworks of culture and living laboratories with many uses in teaching and research. Legislation comparable to what was used to establish the U.S. Library of Congress (http://www.loc.gov) might be needed to encourage cooperation by the owners of the intellectual property.

It will also be necessary to create some curriculum in and around the worlds and tweak some of their parameters so users with different goals are spared having to invest months of effort to gain full access, as they are today. For example, *Pirates of the Burning Sea* allows users who select rare nationalities to advance up the experience ladder twice as fast as others. Both *World of Warcraft* and *Age of Conan* (http://www.ageofconan.com) allow advanced users to create new characters who start their virtual lives already far advanced.

The best of the early virtual worlds are immortal masterworks of culture and living laboratories with many uses in teaching and research.

My avatar in your world (I call him Bill) has held conferences in both *Second Life* and *World of Warcraft*. It would be a simple matter for him to create a high school or college course in either *Pirates of the Burning Sea* or *A Tale in the Desert*. The former would be good not only for courses in history but also for political economy. Quite apart from its Egyptian motif, *A Tale in the Desert* offers challenges in puzzle solving, logic, and the engineering of industrial supply chains. Urban studies could be taught in *The Matrix Online*, and teaching modules incorporating experiments in many social sciences could be added to any of these worlds.

The library where I work in Rivendell is 1,000 years old, and I have trouble imagining all the difficulties you might face if you were to try to build a Digital Library of Virtual Worlds. Yet what a shame it would be if the glorious creativity of the first generations of virtual worlds were truly gone forever.

The first great grand opera, *l'Orfeo*, composed by Claudio Monteverdi in 1607 is still performed today, and anyone may buy a recording for a few dollars. Four hundred years from now, I hope your descendents will still be able to visit me so I can introduce them to Frodo, Bilbo, and Gandalf... and perhaps all go Orc hunting together. C

*__Rumilisoun__ and __Edmund Bainbridge__ are both avatars of __William Sims Bainbridge__; the real Edmund was Bill's great-great-great-great-great-great-grandfather, and Edmund's grandson, Commodore William Bainbridge, was held captive for two years by Barbary pirates.

Future Tense, one of the revolving features on this page, presents stories and essays from the intersection of computational science and technological speculation, their boundaries limited only by our ability to imagine what will and could be.

DOI:10.1145/1859204.1859233

Rumilisoun*

Future Tense
Rebirth of Worlds

Build a digital library of pioneering virtual worlds as a living laboratory of history and social science.

FROM ELROND'S LIBRARY at Rivendell in Middle Earth, I write to you about a serious threat that endangers the connection between our worlds, and possibly the worlds themselves. I am Rumilisoun, an immortal Elf lore-master and historian, responsible for preserving ancient manuscripts, artifacts, and memories. The connection I speak of, between Middle Earth and your Internet, is *Lord of the Rings Online* (http://www.lotro.com), and similar connections exist between our worlds and many others that are in imminent danger. *Lord of the Rings Online* is today quite healthy, though none can predict whether it still will be in 10 years. The history books about Middle Earth written by J.R.R. Tolkien some 60 years ago will still be read in thousands of years, translated into whatever language people use then, and the movies made from them will endure in constantly renewed digital form. Yet many virtual worlds have already died, leaving no copies in libraries.

The Matrix Online closed July 31, 2009, ending forever the possibility of directly experiencing the city depicted in the 1999 movie. An intriguing pace-travel world with an interesting philosophy, *Tabula Rasa*, died February 28, 2009. *The Sims Online* died August 1, 2008. In all three, though the companies operating them calculated they were no longer profitable, they have shown no sign of wanting to put them in the public domain. Contemporary novels in public libraries cut into the profit of trade-book publishers, even though academic publishers depend on libraries, and the owners of virtual worlds have no incentive to give up their property rights. Indeed, allowing a nonprofit organization to operate virtual worlds would compete directly with their commercial counterparts.

Rumilisoun lecturing you at Elrond's Library in *Lord of the Rings Online*.

Transferring virtual worlds to a digital library would entail some cost, in part because they differ so much from one another and need maintenance, and in part because some changes would be needed to make them maximally valuable for researchers, teachers, and students. On the client side, a virtual world consists of a user interface plus graphics files, and on the server side of network-management software and an immense database that reliably describes the moment-by-moment condition of thousands of avatars. Without both sides of the Internet connection and without at least a few hundred inhabitants, a virtual world cannot exist.

Consider the world inhabited by my friend Edmund Bainbridge. Born in New Jersey in 1702, he voyaged at age 18 to the Caribbean for adventure and to try his hand as an English freetrader and shipbuilder in *Pirates of the Burning Sea* (http://www.burningsea.com). He had a marvelous time sailing across a realistic sea in a variety of authentic sailing ships, blasting away with his canon at the French, Spanish, and occasional pirate. He set up timber mills in two forests, one for oak for the hulls of his ships, the other for fir for their masts, operated a sulfur mine because that ingredient for caulking was not available in the auction system, and is able to construct many different craft in his shipyard. He enjoys the cultural life of the most advanced ports, including the recent musical [CONTINUED ON P. 127]

IMAGE COURTESY WILLIAM SIMS BAINBRIDGE

IEEE 7th World Congress on Services (SERVICES 2011)

July 5-10, 2011, Washington DC, USA, http://www.servicescongress.org/2011

Modernization of all vertical services industries including finance, government, media, communication, healthcare, insurance, energy and ...

IEEE 8th International Conference on Services Computing (SCC 2011)

In the modern services and software industry, Services Computing has become a cross-discipline that covers the science and technology of bridging the gap between Business Services and IT Services. The scope of Services Computing covers the whole lifecycle of services innovation research that includes business componentization, services modeling, services creation, services realization, services annotation, services deployment, services discovery, services composition, services delivery, service-to-service collaboration, services monitoring, services optimization, as well as services management. The goal of Services Computing is to enable IT services and computing technology to perform business services more efficiently and effectively. Visit http://conferences.computer.org/scc.

IEEE 9th International Conference on Web Services (ICWS 2011)

As a major implementation technology for modernizing software and services industry, Web services are Internet-based application components published using standard interface description languages and universally available via uniform communication protocols. The program of ICWS 2011 will continue to feature research papers with a wide range of topics focusing on various aspects of implementation and infrastructure of Web-based services. ICWS has been a prime international forum for both researchers and industry practitioners to exchange the latest fundamental advances in the state of the art on Web services. Visit icws.org.

IEEE 4th International Conference on Cloud Computing (CLOUD 2011)

Cloud Computing is becoming a scalable services delivery and consumption platform in the field of Services Computing. The technical foundations of Cloud Computing include Service-Oriented Architecture (SOA) and Virtualizations of hardware and software. The goal of Cloud Computing is to share resources among the cloud service consumers, cloud partners, and cloud vendors in the cloud value chain. Major topics cover Infrastructure Cloud, Software Cloud, Application Cloud, and Business Cloud. Visit http://thecloudcomputing.org.

Sponsored by IEEE Technical Committee on Services Computing (TC-SVC, tab.computer.org/tcsc)

Submission Deadlines

ICWS 2011: 1/31/2011
CLOUD 2011: 1/31/2011
SCC 2011: 2/14/2011
SERVICES 2011: 2/14/2011

Contact: Liang-Jie Zhang (LJ) at zhanglj@ieee.org (Steering Committee Chair)

IEEE TRANSACTIONS ON SERVICES COMPUTING

CSCW 2011
BUILDING BRIDGES

The 2011 ACM Conference on
Computer Supported Cooperative Work
March 19–23, 2011 · Hangzhou, China

Register Online Today
Early registration discount until January 7, 2011

Reserve Your Hotel Room
Discounted reservations at the conference hotel, the **Hyatt Regency Hangzhou**

CSCW is an international and interdisciplinary conference that has a special focus on how technology intersects with social practices. Join us in **Building Bridges** by connecting with social and technical researchers at the conference and interacting with research communities from around the world.

Conference Co-chairs
Pamela Hinds, *Stanford University*
John C. Tang, *Microsoft Research*
Jian Wang, *Alibaba Group*

Keynote Speakers

Mr. Jack Ma
CEO and Chairman, Alibaba Group

Dr. Genevieve Bell
Director of Interaction & Experience Research, Intel Corporation

Tutorials
Mobile User Experience Essentials
Supporting Multilingual Communities
Intro to CSCW Research
... and more!

Find CSCW 2011 Online
www.cscw2011.org
www.facebook.com/cscw2011
www.twitter.com/cscw2011

Sponsored By
acm
SIGCHI
special interest group computer human interaction