DOI:10.1145/2500892

Keywords in the ACM Digital Library and IEEE Xplore digital library and in NSF grants anticipate future CS research.

BY APIRAK HOONLOR, BOLESLAW K. SZYMANSKI, AND MOHAMMED J. ZAKI

Trends in Computer Science Research

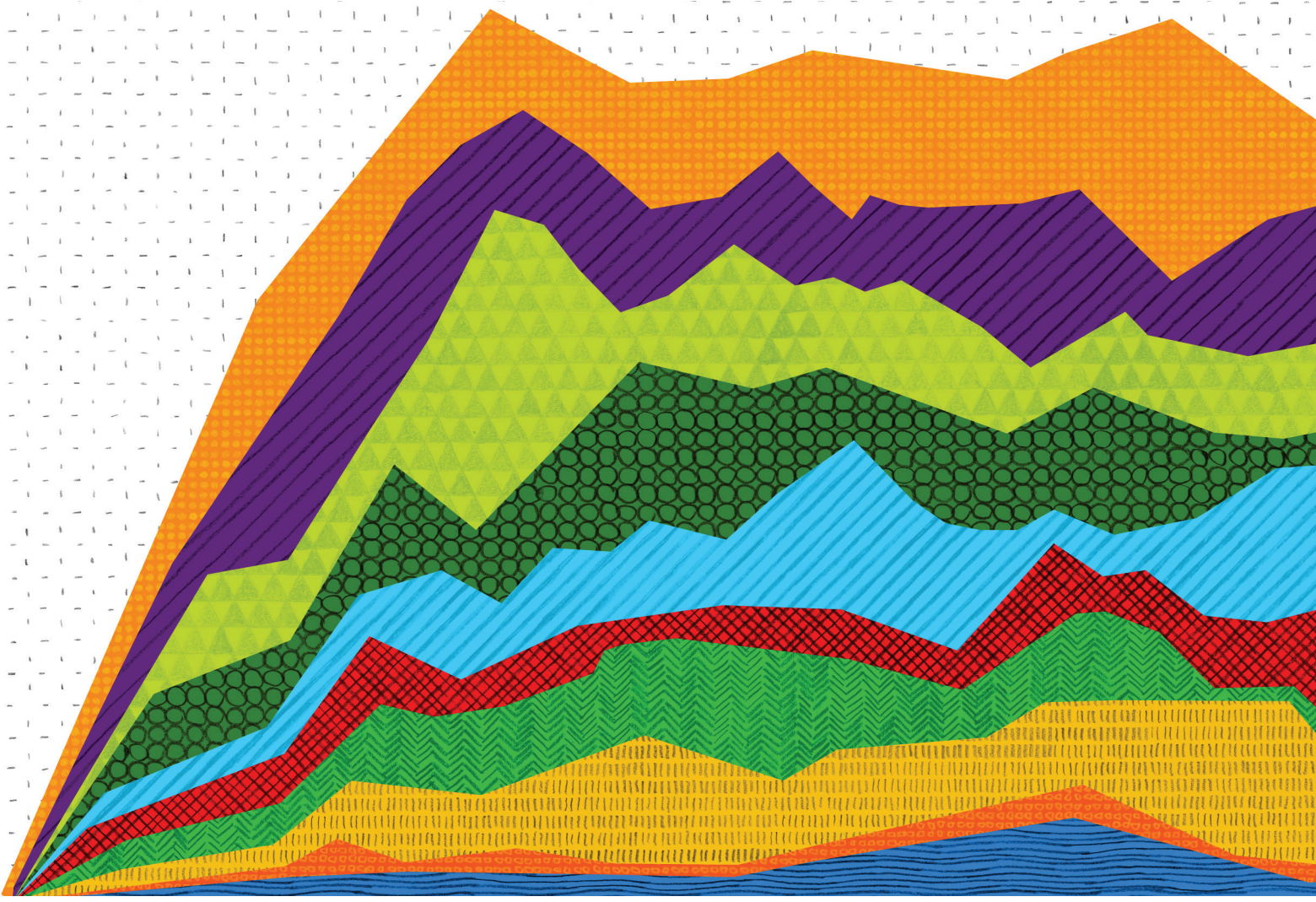
COMPUTER SCIENCE IS an expanding research field driven by emerging application domains and improving hardware and software that eliminate old bottlenecks even as they create new challenges and opportunities for CS research. Accordingly, the number of research papers published in CS conferences and journals has been increasing rapidly for the past two decades. With growing emphasis on externally funded research in most universities, scientific research is increasingly influenced by funding opportunities. Although many funded programs are developed in close collaboration with leading researchers, we aim here to identify more precisely relationships between funding and publications related to new topics.

Trend analysis has long been researched and applied to many types of datasets, from medical[17] to weather[15] to stock markets.[5] Many publications track research trends, analyze the impact of a particular paper on the development of a field or topic, and study the relationships between different research fields. The Web of Science[22] has collected data since 1900 on nearly 50 million publications in multiple scientific disciplines and analyzed it at various levels of detail by looking at the overall trends and patterns of emerging fields of research and the influence of individual papers on related research areas. Over the past decade, besides the Web of Science, studies have also investigated the overlap and evolution of social communities around a field or a topic. Rosvall and Bergstrom[18,19] explored methods and visualizations for scientific research and analyzed the impact of each research area quantified by the collective cross-disciplinary citations of each paper. Porter and Rafols[16] analyzed citation information to find evidence of collaboration across fields in scientific research. Other examples are network models for studying the structure of the social science collaboration network[13] and women's authorship of CS publications in the ACM digital library.[3]

Several studies have focused on challenges, directions, and landscapes in specific CS fields[2,7] and on

» key insights

- **A burst of new keywords in grants generally precedes their burst in publications; less than one-third of new keywords burst in publications first, reflecting the importance of funding for success of new CS fields.**
- **A typical scientist's research focus changes in roughly a 10-year cycle and often includes a once-in-a-career shift, likely in response to evolving technology creating new CS fields.**
- **CS continues to experience continuous and fundamental transformation; for example, in the past two decades, new topics arose within the Internet research cluster, while some previously popular topics (such as mathematical foundations) decayed.**



specific CS topics.[8,21] In this article, we are interested in learning about the evolution of CS research. We collected data from 1990 to 2010 on proposals for grants supported by the U.S. National Science Foundation[14] and on CS publications in the ACM Digital Library[1] and IEEE Xplore digital library.[11] We analyzed research communities, research trends, and relations between awarded grants and changes in communities and trends, as well as between research topics. We found if an uncommonly high frequency of a specific topic is included in publications, the funding for the topic usually increases. We also analyzed CS researchers and communities, finding only a small fraction of authors attribute their work to the same research area for a long time, reflecting an emphasis on novelty (new keywords) and frequent changes in academic research teams (a stable core of faculty and turnover of students and postdocs). Finally, our work highlights the dynamic CS research landscape, with its focus constantly moving to new challenges due to new technological developments. CS is an atypical academic discipline in that its universe is evolving so quickly, at a speed unprecedented even for engineering. Naturally, researchers follow the evolution of their artifacts by adjusting their research interests. We attempt to capture that vibrant coevolution here.

We used the ACM, IEEE, and NSF datasets from which we collected data on publications from 1990 to 2010.[a] For the ACM dataset, we extracted the number of papers listed in top categories of the 1998 ACM Computing Classification System, or CCS.[b] The ACM dataset included authors, title, abstract, year published, publication venue, author-defined keywords, and ACM classification categories for each of the 116,003 articles. We used the ACM CCS and author-defined keywords to respectively study the broader and static versus the finer and dynamic views of the CS landscape and trends. In another analysis, we used only the author-defined keywords to identify the relationships between researchers, yielding smaller research groups than if we had used just ACM CCS alone.

The IEEE Xplore dataset included similar information but lacked a topic classification like the ACM CCS. Instead, we used 408 research topics included in 16 Wikipedia articles on CS research areas identified in the main Wikipedia CS article[23] to classify 458,395 papers in the IEEE dataset. For the NSF dataset, we retrieved titles, start dates, and abstracts of 21,687 funded grant proposals.

For the ACM and IEEE datasets, we created two data indexes—authors and their publication venues and papers and their keywords/topics—finding, in the analyzed period, the number of publications grew approximately 11% yearly over those 20 years. To create research topic networks, we made each

a NSF records before 1990 were incomplete (such as lacking abstracts), but only 10% of publications in the ACM and IEEE datasets were published before 1990, so our time range covers nearly all publications in those datasets.

b We excluded the "general literature" category because CCS (http://www.acm.org/about/class/1998/) includes too many non-research topics (such as biography and reference).

ILLUSTRATION BY ALICIA KUBISTA/ANDRIJ BORYS ASSOCIATES

topic a node and connected two nodes with a weighted edge representing the number of abstracts that mention both adjacent topics.

Using sequence mining,[24] network extraction and visualization,[18] bursty words detection,[12] clustering with bursty keywords,[c,10] and network evolution,[6] we investigated changes over time in the CS research landscape, interaction of CS research communities, similarities and dissimilarities between research topics, and the impact of funding on publications and vice versa.

Results and Discussion

Landscape of CS research. We looked at the evolution of the CS research landscape 1990–2010 (see Figure 1). Many ACM records 2009–2010 (collected during the spring of 2011) did not have ACM classification categories and thus were excluded from our study, causing the drop off of records in the last two years in Figure 1. (There is no such drop in Figure 4 because every record included a publication venue.) For ACM, except for 2009 and 2010, publications in each category increased year over year, but after 1994 the fraction of publications in the "mathematics of computing" category shrank considerably. The author-defined keywords contributing to the drop were control theory and logic. We attribute the drop to a shift of focus from general issues to challenges specific to an area with which such publications are increasingly associated. For the remaining categories, the fastest growing were publications in information systems. The most frequently used author-defined keywords were Internet-related (such as XML, Internet, Web services, and semantic Web). Likewise, the IEEE dataset showed the fastest-growing research area was information science and information retrieval.

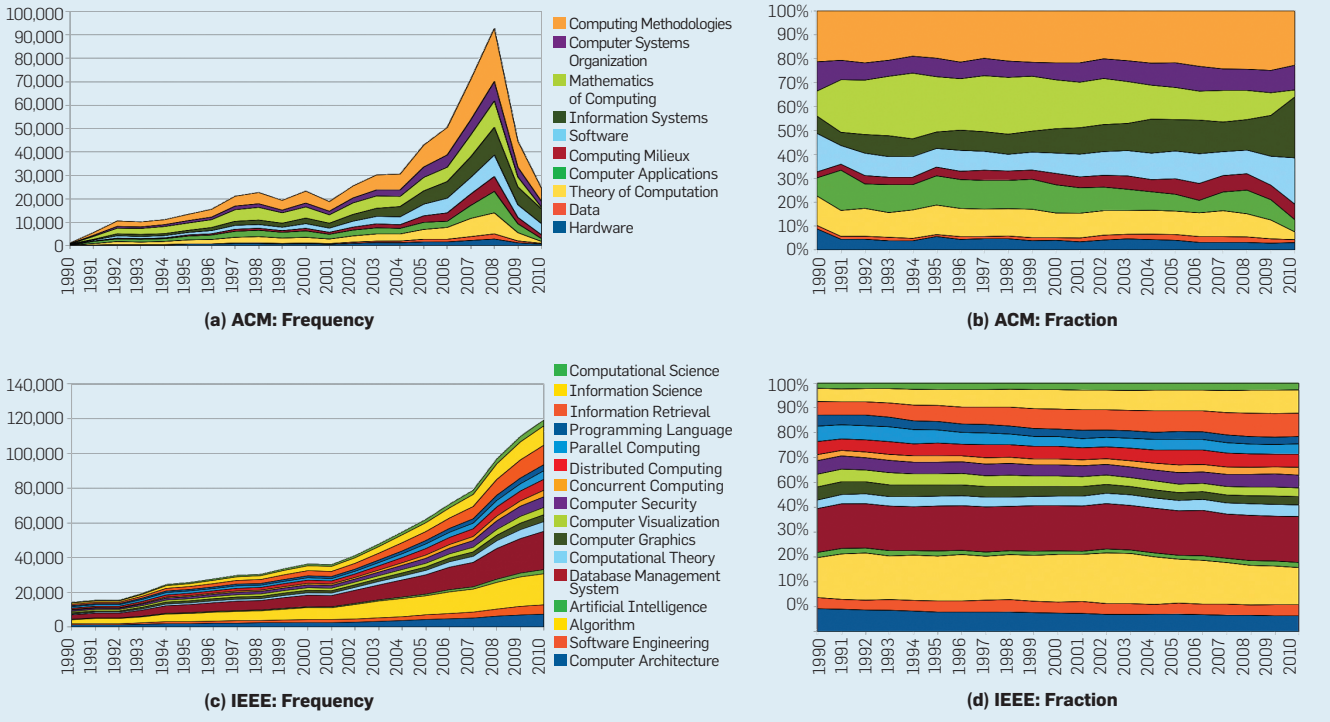
To better see the impact of information systems, we extracted the top 25 research topics from ACM and IEEE and quantified the results in two ways: document frequency (DF) and term frequency inverse document frequency (TFIDF). For term/keyword k, DF is the number of documents including it; TFIDF is the sum of the weights over all documents, where the weight of k in document d is defined as

$$\frac{n_{k,d}}{\sum_{w \in d} n_{w,d}} \cdot \log \frac{|D|}{|j : k \in d_j|}$$

where $|D|$ is the number of documents and $n_{k,d}$ is the number of times k appears in d; see Hoonlor et al.[9] for detailed results. Most publications in collaboration, data mining, information retrieval, machine learning, privacy, and XML appeared 2000–2010 and showed noteworthy trends in CS research. The terms Internet and World Wide Web did not appear in any publication until 1995, but the related topics were present since early 1990. During the period 1990–1997, 376 NSF grants and nine IEEE papers mentioned NSFNET in their abstracts, but only two ACM papers included it as a keyword. Other terms (such as net, prodigy, point-to-point, and internetworking) also appeared in the NSF dataset before 1995. More-

c The term "bursty keywords" in this context refers to keywords appearing with uncommonly high frequency during some intervals; such intervals may include multiple spikes of a keyword's frequency.

Figure 1. Two views of CS research, 1990–2010, based on the ACM (a) and (b) and IEEE (c) and (d) datasets; frequency is number of publications on each topic each year; fraction is the percentage of publications on each topic each year.



over, prodigy was bursty over the period 1991–1992 and TCP/IP over the period 1990–1993.

TFIDF and DF values showed the rise of information system contributed to the general interest in data mining, information retrieval, and Web-related topics, whereas mathematics of computing continued to decrease year over year during the same period, with logic and control theories contributing most to the decline.

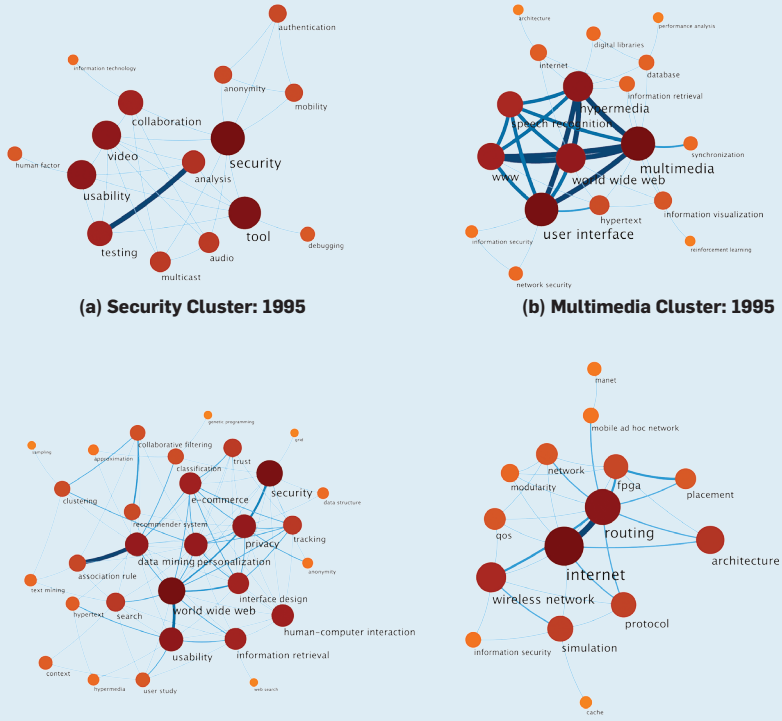
Figure 2 includes the research topic subnetworks culled from the ACM dataset by the Map Generator software package[4] for the security and the multimedia subnetworks found in 1995 and for the World Wide Web and the Internet subnetworks found in 2001. In 1995, Web was used as a keyword associated mostly with multimedia and information visualization, whereas information retrieval was used mostly with Internet. However, by the early 2000s, Web was used mostly with data mining and information retrieval, while Internet was used mostly with network, protocol, and routing. Since 2005, privacy and security have become important in the Web context, while semantic Web, Web 2.0, Web service, and XML became major Internet topics.

Bursty-period analysis. To evaluate the influence of research funding on publications and vice versa, we extracted from ACM,[d] IEEE, and National Science Foundation datasets the bursty periods of author-defined keywords based on the burstiness score for a time period[12] defined as

$$Burst(w,t) = \left(\frac{|d_t : w \in d_t|}{|d : w \in d|} - \frac{1}{T} \right)$$

where w is the keyword/topic of interest, t is a time period, d_t is a document created during time t, d is any document, and T is the total time over which all documents were created. The burstiness score measures how often w is in t compared to its occurrence in T. A positive score implies w appears more often during the "bursty period" t than over the total time T. We recovered the maximal segments of burstiness scores in the sequence of documents using the linear-time maximum sum algorithm;[20] each segment with positive score corresponds to a bursty period.

Figure 2. Research clusters, or subnetworks, in 1995 and 2001; edge thickness represents strength of interaction.



(a) Security Cluster: 1995

(b) Multimedia Cluster: 1995

(c) World Wide Web Cluster: 2001

(d) Internet Cluster: 2001

For each pair of datasets, we analyzed in which one a keyword's bursty period begins first and then how long it takes for the keyword to become bursty in the other. For keywords with more than one bursty period, we also looked at their burstiness score in each bursty period, then tabulated the percentage of cases in which the later burstiness scores increased, decreased, or was unchanged.

For an ACM-NSF pair, if a keyword became bursty in ACM data first, it became bursty in NSF 2.4 years later on average; in the reverse case, the average delay was 4.8 years. The longer delay shows if NSF initiates a new area, the increase in publications is delayed by the time researchers need to obtain grants and start research leading to publication. If the keywords were bursty in both datasets, in 75% of such cases the keyword became bursty in the NSF dataset before it became bursty in the ACM dataset, showing NSF funding often increases interest in the supported areas. For another 16% of cases, it was reverse; examples of bursts appearing first in the NSF dataset are data mining and search engine, becoming bursty in 1999 for NSF and in 2000 for ACM. The reverse included bioinformatics (2003 in ACM and 2004 in NSF) and semantic Web (2004 in ACM and 2006 in NSF).

For an IEEE-NSF pair, a keyword first bursty in IEEE became bursty in NSF 3.4 years later on average; in the reverse case, the average delay was 5.7 years. The difference between these two delays and its causes are the same as in the ACM dataset. Yet both delays are one year longer than in the ACM-NSF pair, resulting, we conjecture, from a larger ratio of computer engineering topics in IEEE than in ACM and presumably to a larger fraction of support for IEEE publications from non-NSF sources.

In two-thirds of the cases where a keyword was bursty in both the NSF and the IEEE datasets, it became bursty in the NSF dataset first, consistent, again, with the ACM dataset. The other cases were evenly split between the reverse and the concurrent appearance of burstiness in both datasets. Only Internet (in 2000) and telecommunications (in 1995) became bursty at the same

d Since only ACM records are classified through CCS, we do not use that classification here.

time in both datasets. Keywords bursty first in the IEEE dataset included real-time database (1994 versus 1999 for NSF), procedural programming (1992 versus 1993), and neuro-biological (1996 versus 2001). Peer-to-peer network was bursty in IEEE 2003–2010 but never in the NSF dataset, possibly indicating the corresponding challenges were funded mostly by non-NSF sources; see Hoonlor et al.[9] for more detailed bursty-period comparisons.

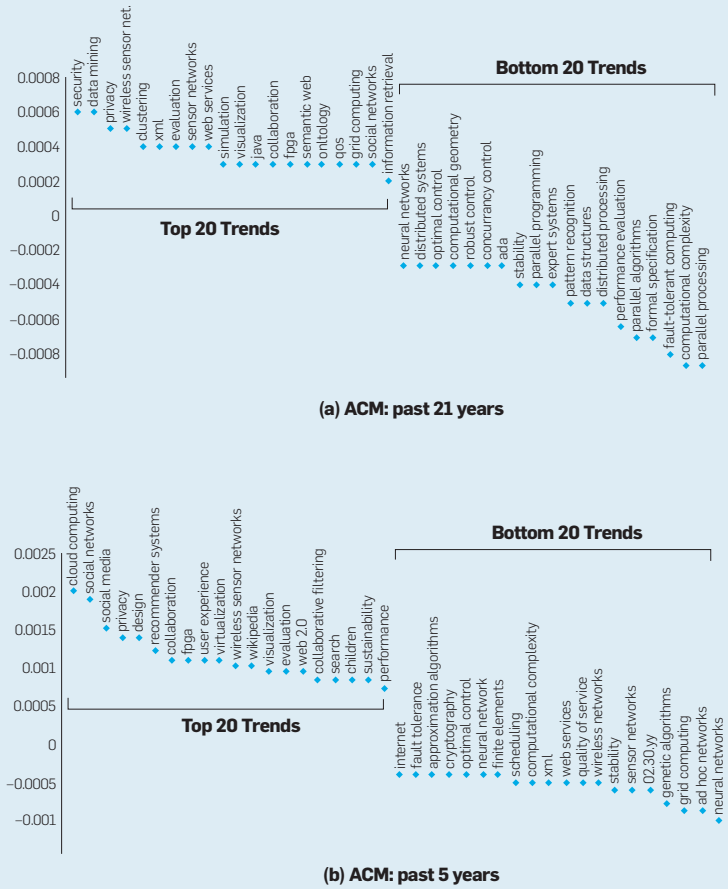
We also analyzed the NSF dataset versus ACM and IEEE datasets and vice versa. For each such pair and each year 1990–2010, we searched for the year in which the number of entries changed compared to any of the previous four years in the first database. For each change, we searched in the other dataset for a change in any of the next four years. The relative change values ranged from −0.5 to 0.5, which we grouped into bins of size 0.1. We counted the frequency of the change in one dataset followed by a change in the other.

For the NSF dataset versus either the ACM or the IEEE dataset, a 10% or greater increase in the number of NSF grants awarded for a given topic from the previous few years was followed by an increase (with 75% probability) in the number of published papers on the topic of at least 10% in the next three years and 20% in the next four years. Topics with such an increase included data mining, information extraction, and wireless network. On the other hand, an increase of 10% in the number of published papers in a given topic in the ACM dataset was followed with 75% probability of an increase (usually less than 10%) in the number of NSF grants awarded on the same topic; examples were e-government, groupware, and knowledge management.

For a keyword with multiple bursty periods in the NSF dataset, the following bursty period had a higher/lower/equal burstiness score in 37%/51%/12% of the cases. For the IEEE dataset, it was 29%/64%/7%, respectively, and for the ACM dataset, it was 12%/85%/4%. However, for interleaved or overlapped bursty periods in the NSF and IEEE datasets, if the bursty period appeared first in the IEEE dataset, the following NSF bursty period had a higher/lower/equal burstiness score in 31%/22%/47% of the cases. In the reverse case, it was 36%/10%/55%. The same analysis of the ACM and NSF datasets showed the following NSF bursty period had higher/lower/equal burstiness score for 38%/14%/48% of the cases; in the reverse case, for the following ACM bursty period, the numbers were 8%/8%/84%.

The reason for a large percentage of equal burstiness scores is that a bursty period in one dataset was often a subset of the bursty period in another. Burstiness scores tend to decrease in the periods following a bursty period in the NSF dataset. Since novelty is prized so highly in publications, authors tend to stress new aspects of their work in abstracts and keywords, contributing to the observed pattern. Yet during an NSF burstiness period, publication burstiness scores were more likely to increase than decrease, confirming sustained NSF funding is essential for maintaining interest in a given topic.

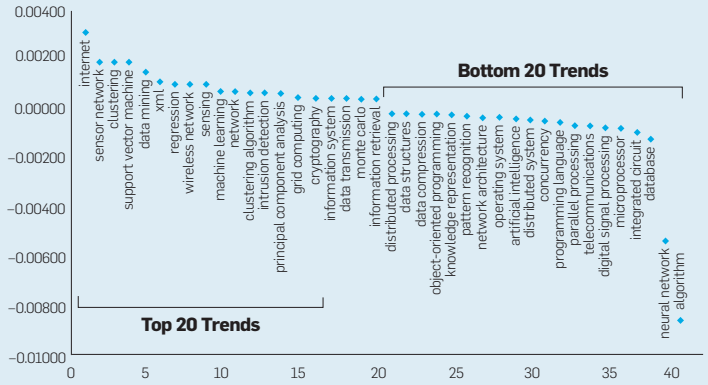
Figures 3a and 3b. Top 20 and bottom 20 trends 1990–2010 and 2006–2010 from the ACM and IEEE datasets.



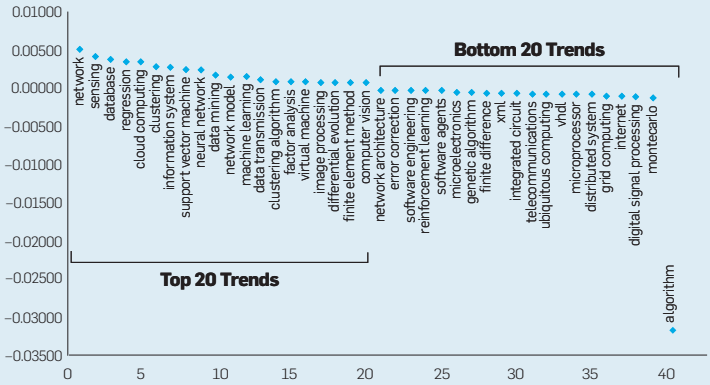
Further analysis identified keywords associated with each bursty period that burst together; for example, wireless sensor networks are temporally related to simulation, security, and clustering in the order of bursty periods. This order corresponds to the temporal evolution of the area that initially focused on simulation of networks, then on security, and finally on clustering algorithms. The analysis also revealed data mining is more broadly used than information retrieval. The former is used with computational science, Web mining, time series mining, and security; the latter is used mainly with Web-related topics. Text mining is temporally related to both information retrieval and data mining.

Multiple bursty periods for a keyword include interesting temporally correlated terms. For example, there were three bursty periods for the keyword "scheduling": 1990–1991, 1999, and 2001–2006. In the bursty period of 1999, scheduling correlated (listed

Figures 3c and 3d. Top 20 and bottom 20 trends 1990–2010 and 2006–2010 from the ACM and IEEE datasets.



(c) IEEE: past 21 years



(d) IEEE: past 5 years

in the order of burstiness ranking) with genetic algorithms, parallel processing, performance evaluation, embedded systems, approximation algorithm, multimedia, quality of service, optimization, and heuristics. In the period 2001–2006, such keywords, listed in the same order, were approximation algorithms, multimedia, online algorithms, real time, embedded systems, fairness, multiprocessor, quality of service, and genetic algorithms. Hence, initially, both real-time systems and parallel processing were related to scheduling, later expanding to genetic algorithms and embedded systems. In the last few years of its bursty periods, scheduling also correlated with multimedia, online algorithm, and fairness. An alternative look at such links can be done through the co-referenced document frequency instead of the burstiness score.[9]

Trend analysis. Here, we analyze research trends through the linear regression trend line and changing popularity of topics based on the fraction of papers including a given keyword in each year. We generated a trend line for each keyword fraction and used its slope for ranking. We fit the trend lines to data from the preceding two to six years in order to predict keyword fractions for the following year.

In all datasets, we observed that if a trend based on two years of data had a positive slope, or the fraction of publications increased from the previous year to the current year, then the subsequent year fraction declined. We also used the trend line based on the NSF dataset to predict fractions for the following year in the ACM and IEEE datasets. The results show the trend line is a poor predictor, as is using ACM and IEEE trends to predict the number of grants awarded by NSF; the accuracy of all these models was less than 50%.

Figure 3 includes the top 20 up and top 20 down trends for the period 1990–2010 and for the period 2006–2010 for ACM and IEEE. Unlike the ACM dataset, the IEEE dataset did not show a significant decrease between the top and the bottom trends because research topics appear in the abstracts longer than do author-defined keywords. Moreover, we used the CS conferences listed in Wikipedia[23] to categorize each paper in IEEE and ACM; see Hoonlor et al.[9] for the complete list of categories. We could not statistically compare the growth in different areas due to vast differences in the number of conferences in each field and number of papers published in each conference. Nevertheless, Figure 4 indicates growth of approximately 11% in publications for most CS topics year over year, with each topic representing a set of CS conferences;[e] when a conference covered two topics, then papers published in the conference were indexed in both topics.

Network of CS research. Since we looked back over the period 1990–2010, we were also able to monitor when connections between two fields occurred or changed. We extracted two sets of keywords: those never appearing together in the same article and those appearing together in at least some specified number of articles each year. For the IEEE dataset, keeping algorithm as a node greatly reduced the degree of separation between other research topics and created a central node, or one with the highest total weight of its edges, dominating other research topics. We performed the network analysis on the algorithm topic first, then removed the algorithm node from the network because the term was used in almost every CS research paper to describe how data is processed.

During the period 1990–2010, algorithm, database, and neural network were the most frequent CS research topics; 311 other CS research topics were also mentioned, along with algorithm at least once, with 78 of them persistent; that is, they co-appeared with algorithm every year from 1990

e In contrast, Figure 1 uses ACM classification and IEEE Xplore keywords; even ACM records missing ACM classification terms are represented here since each record includes publication information.

to 2010. Of 408 CS research topics, 286 were mentioned with database, but only 32 were persistent topics.[f] Meanwhile, 254 topics appeared with neural network, with only 39 persistent. Besides the three most frequent topics, 11 others had persistent connections with multiple research topics every year 1990–2010, including programming language, artificial intelligence, clustering, image processing, computer vision, network, distributed system, pattern recognition, robotics, software engineering, and integrated circuit. Also during 1990–2010, 87 other research topics, including image analysis, data transmission, and operating system, were linked with up to three of the mentioned 14 topics.

In ACM networks using author-defined keywords, no persistent link appeared 1990–2010, confirming our earlier observation that while a certain research topic may be important enough to be mentioned in an article's abstract, it may not represent the article's key research contribution. Another example of lack of link persistence is the neural network node in both IEEE and ACM networks. In the former, neural network was a central node in almost every year. Yet in ACM networks, it never achieved this status. Lack of link persistence is also evident for algorithm and database topics. In the early 1990s, user interface, scheduling, and multimedia were associated with many CS research fields. However, in the late 1990s, interest shifted to the Web, information retrieval, and computer-supported cooperative work. Throughout the 2000s, the areas most connected to others were design, usability, and security; the mid-2000s saw strong interest in sensor network and later in wireless sensor network.

We performed clustering on the yearly network of keywords in the ACM dataset in which a keyword can appear in multiple clusters; using the clusters, we measured the similarity between keywords k and a as

$$\frac{\text{Number of clusters with } a \text{ and } k}{\text{Number of cluster with } a}$$

An increase of 10% in the number of published papers in a given topic in the ACM dataset was followed with 75% probability of an increase (usually less than 10%) in the number of NSF grants awarded on the same topic.

We found a list of terms clustered together with network connectivity in the period 2006–2010 though not connected in at least 1% of the documents.[24] We examined the top 10 frequently used keywords at various degrees of separation; see Hoonlor et al.[9] for results. During the period 2006–2010, simulation was clustered with many keywords in database research, including integration, data warehouse, and relational database, although they were either not used or used only rarely by authors to describe their research in simulation. Simulation was instead clustered with information retrieval, feature selection, and filtering, as well as with other topics related to data mining, machine learning, and artificial intelligence, but was not used directly to describe the same research project often enough. Data mining was rarely used to describe research related to mobile networks and its related research topics.

CS researchers. We used the cSpade sequence mining algorithm[24] to analyze sequences of publications in the same major research category by the same author, requiring at most a one-year gap in publication dates and appearance in at least 1% of documents. We recorded the maximum length of publication sequences in the same category and measured the percentage change in the number of publications of a given author after the first year in each category. From all the authors whose publications were in the same categories, we calculated the half-life time (it took for the number of authors who continued publishing papers in the category to reduce by half). For the first analysis, we used the ACM CCS to identify major research categories, then performed the same analysis using the list of conferences under six CS categories, finding the rates of publication growth differed in each category; see Hoonlor et al.[9] for details.

We found a relatively short half-life time, as well as a significant first-year drop rate, especially for computer application, computing milieu, and data keywords, indicating the authors in these categories were either only briefly involved in multiple research topics or

f Top five persistent topics in the database research cluster were relational database, distributed database, database management, query language, and database design; for the neural network research cluster, they were pattern recognition, regression, supervised learning, reinforcement learning, and robotics.

only briefly collaborated with someone else from these categories. Researchers in computer systems organization, computing methodologies, and information systems tended to stay active in these categories for a longer time. Moreover, we found it difficult for researchers to publish in artificial intelligence and programming language year after year, unlike in, say, human-computer interaction. Researchers in human-computer interaction remain active the longest, followed by researchers in computer architecture; see Hoonlor et al.[9] for details.

Note while researchers can continue to publish in one area for a long time, the area itself evolves and may cover different topics during different time periods; for example, human-computer interaction focused mainly on interaction design, visual design, and computer-supported cooperative work in the 1990s and augmented reality, computer vision, human factors, and ubiquitous computing in the early 2000s, finally shifting to social media, learning, computer-mediated communication, and tangible user interface in the late 2000s.

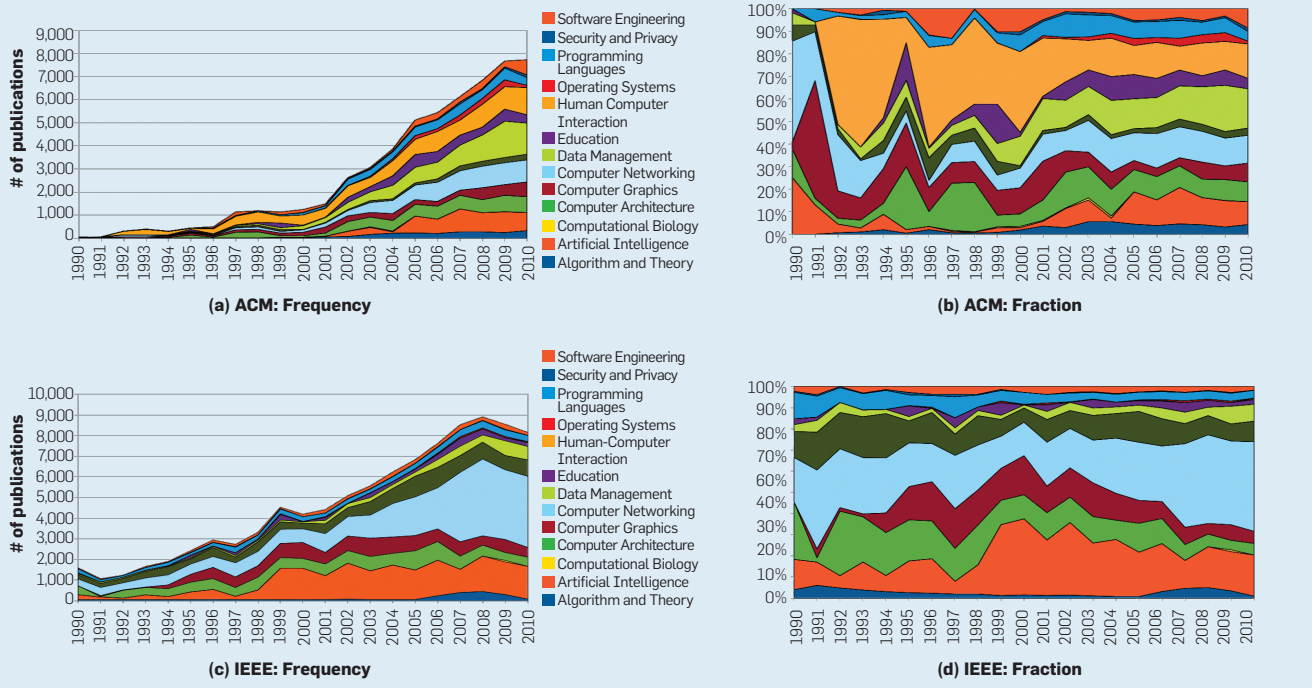
Investigating further, we selected four prominent CS researchers, analyzed their publications, and discussed the results with them. Jack Dongarra of the University of Tennessee, Knoxville, is renowned for developing high-performance linear algebra software packages and systems, though his interests have evolved over time. In the 1980s, for example, he focused on parallel algorithms for linear equation routines and linear algebra subprograms. In the early 1990s, he focused on parallel solutions for eigenvalue problems and numerical software libraries for high-performance systems. From the late 1990s to the 2000s, he focused on high-performance linear algebra packages for multicore systems. More recently, he has also focused on performance of grid computing. Overall, his research interests have evolved continuously in response to challenges created by new computer technologies.

Another researcher in this area, Francine Berman of Rensselaer Polytechnic Institute, Troy, NY, characterized her work in 1980s as "top-down mathematical modeling" of mapping and scheduling problems. In the early 1990s, her papers used such keywords as data-driven, performance, and algorithms. From the late 1990s to the mid-2000s, she focused on grid computing from a "bottom up" perspective: application-level scheduling/rescheduling, job distribution, and performance. She described this evolution as a broadening and branching approach. Since 2003, she has made a major shift to large-scale cyberinfrastructure and data preservation.[g]

In the early 1990s, George Cybenko of Dartmouth College, Hanover, NH, studied high-performance computing and classification by neural networks. In the late 1990s, he shifted to mobile agents, mobile networks, and simulations. In the early 2000s, he focused on target tracking, analyzing data, and extracting information from the Web and from wireless networks. Since 2002, he has investigated privacy and security issues, including cybersecurity, saying he investigates each subject "in five-year (more or less) phases" then "discovers an open field often related to previous work." One exception was a major shift in 1992 when moving from one university to another.

g Cyberinfrastructure and data preservation did not show up as her keywords because the relevant publications were too new to be in our databases.

Figure 4. Landscape of CS research fields, based on conferences 1990–2010, for the ACM and IEEE datasets, including raw numbers (frequencies) and percentage of publications for each keyword each year.

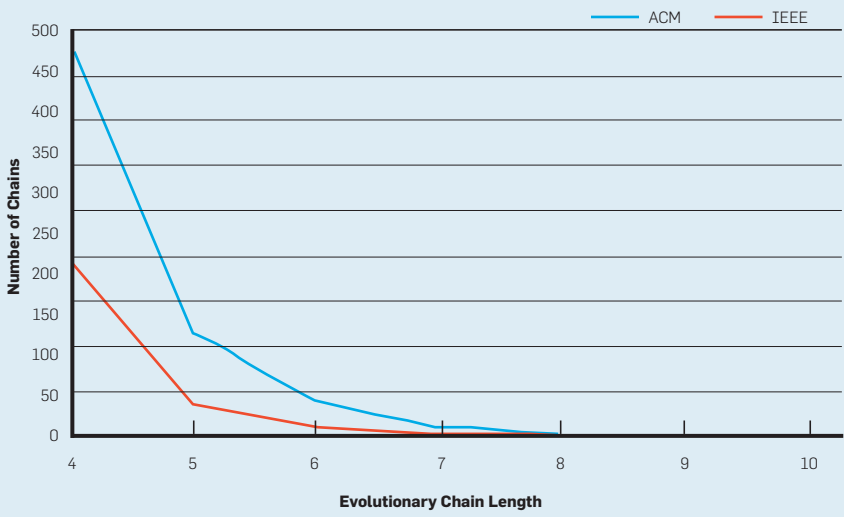


As a final example, James A. Hendler of Rensselaer Polytechnic Institute, Troy, NY, has worked in artificial intelligence since the late 1980s. His major shift was from planning and Web intelligence to the semantic Web. From the late 1980s to early 1990s, his work focused on planning in artificial intelligence and later on agents, real-time systems, and Web technology. In the 2000s, he focused mainly on the semantic Web and for the past few years also on large data and social networks.

Overall, CS faculty research interests typically evolve every five to 10 years by broadening their scope and branching into new applications, as well as responding to technological innovation. Less frequently, perhaps once in a career, there is a major shift to a new area.

Communities of CS researchers. Using the framework for analyzing the evolution of social communities developed by Goldberg et al.,[6] we tracked the evolution of CS researcher communities by searching for overlapping communities over consecutive time periods. We used the networks of authors represented as a bipartite graph in which each node representing a paper has edges to all nodes representing the paper's authors (see the table here). Figure 5 plots the number of communities that survived from one year to the next in the ACM and the IEEE datasets. The table lists average evolutionary chain length, average cluster size, average size of intersections of two to four consecutive clusters, and average relative density.[h] We found the recovered clusters had high average relative density of 0.8 for both datasets. The average length of the evolutionary chain was 4.5 years, while approximately two core researchers were associated with each cluster. This finding was consistent with the typical university team consisting of one or two stable faculty and three to five graduate students and postdocs joining and leaving continuously. Every four years or so, only a few stable researchers typically remained from an original research group.

Figure 5. Distribution of the length of evolutionary chains showing number of years a slowly evolving research community remains continuously active based on the ACM and IEEE datasets.



Evolution of research communities in terms of average size of a research group and number of years it was active based on the ACM and IEEE datasets.

Dataset	Average Value of	
ACM	Chain length	4.48
	Cluster size	6.1
	Intersection of two consecutive clusters	3.45
	Intersection of three consecutive clusters	2.51
	Intersection of four consecutive clusters	2.0
	Density	0.84
IEEE	Chain length	4.39
	Cluster size	5.53
	Intersection of two consecutive clusters	3.17
	Intersection of three consecutive clusters	2.36
	Intersection of four consecutive clusters	1.90
	Density	0.80

Conclusion

Most CS publications mention keyword algorithms, which is not surprising, and most abstracts mention one or more topics related to database, neural networks, and Internet. Our investigation also found the Web has become an attractive source of data and application testbeds, pulling in various researchers working on data mining, information retrieval, cloud computing, and networks. Most research related to the Internet has been done since 2000, even though the concept was introduced shortly after standardization of the TCP/IP protocol suite in early 1980s. Web pages evolved from simple text written in mark-up languages (such as HTML and XML) to the semantic Web, where ontologies are a key component for information retrieval by both humans and machines.

While overall trends provide a clear picture of the direction each topic is taking, the fraction of publications on each topic oscillates from year to year to the point the direction of change in one year consistently reversed in the subsequent year. The same pattern was reflected in the number of grants awarded for each topic each year. Since novelty is prized in publications and grant applications, authors tend to stress novel aspects of their work in abstracts and keywords, contributing to the observed pattern. We also found strong evidence of money preceding research; that is, if

h Average relative density is computed by dividing the combined weight of all edges with both endpoints in a cluster by the combined weight of all edges with at least one endpoint in the cluster.

a research topic bursts in terms of NSF grants first, it is likely to burst in publications within a few years. The opposite pattern is at least twice less frequent. Hence, we conclude while funding is not necessary in the initial growth in a CS research topic, it is essential for maintaining research momentum and researcher interest.

Most authors manage to publish at most once a year in a particular research field. Moreover, authors tend to publish in the same major research category for at most only a few years. Only a fraction of them continues to publish in the same field year after year for a long time. This agrees well with the model of an academic research team in which permanent faculty represent only a small fraction of the overall team of faculty, students, and postdocs, with the latter routinely changing topics after leaving a team. Moreover, a faculty member is often active in more than one area. Finally, since novelty is prized, authors tend to pursue new directions in their research, as reflected in an article's abstract and keywords, further contributing to the observed pattern.

Acknowledgment

The authors thank Francine Berman and James A. Hendler of Rensselaer Polytechnic Institute, George Cybenko of Dartmouth, and Jack Dongarra of the University of Tennessee, Knoxville, for discussions on the evolution of their research interests. The authors also thank Katie Bahran for help editing the article. The research was sponsored by the Army Research Laboratory and accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions in this article are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. government. The U.S. government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation here on.

During an NSF burstiness period, publication burstiness scores were more likely to increase than decrease, confirming that sustained NSF funding is essential for maintaining interest in a given topic.

References

1. ACM Digital Library; http://dl.acm.org
2. Agrawal, R., Ailamaki, A., Bernstein, P.A. et al. The Claremont report on database research. *Commun. ACM 52*, 6 (June 2009), 56–65.
3. Cohoon, J.M., Nigai, S., and Kaye, J. Gender and computing in conference papers. *Commun. ACM 54*, 8 (Aug. 2011), 72–80.
4. Edler, D. and M. Rosvall, M. Map Generator software package. MapEquation, Umeå, Sweden, 2010; http://www.mapequation.org
5. Fama, E.F. The behavior of stock-market prices. *Journal of Business 38*, 1 (Jan. 1965), 34–105.
6. Goldberg, M., Magdon-Ismail, M., Nambirajan, S., and Thompson, J. Tracking and predicting evolution of social communities. In *Proceedings of the Third IEEE International Conference on Social Computing* (Boston, Oct. 9–11). IEEE, 2011, 780–783.
7. Hall, M., Padua, D., and Pingali, K. Compiler research: The next 50 years. *Commun. ACM 52*, 2 (Feb. 2009), 60–67.
8. Hendler, J. and Berners-Lee, T. From the Semantic Web to social machines: A research challenge for AI on the World Wide Web. *Artificial Intelligence 174*, 2 (Feb. 2010), 156–161.
9. Hoonlor, A., Szymanski, B.K., Zaki, M.J., and Thompson, J. *An Evolution of Computer Science Research.* RPI Technical Report 12-01, Rensselaer Polytechnic Institute, Troy, NY, 2012; http://www.cs.rpi.edu/research/tr.html
10. Hoonlor, A., Szymanski, B.K., Zaki, M., and Chaoji, V. Document clustering with bursty information. *Computing and Informatics 31*, 6 (Dec. 2012), 1533–1555.
11. IEEE Xplore; http://ieeexplore.ieee.org/Xplore/
12. Lappas, T., Arai, B., Platakis, M., Kotsakos, D., and Gunopulos, D. On burstiness-aware search for document sequences. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Paris, France, June 28–July 1). ACM Press, New York, 2009, 477-486.
13. Moody, J. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review 69*, 2 (Apr. 2004), 213–238.
14. National Science Foundation Award Search; http://www.nsf.gov/awardsearch/
15. Neilson, R.P. High-resolution climatic analysis and southwest biogeography. *Science 232*, 4746 (Apr. 1986), 27–34.
16. Porter, A.L. and Rafols, I. Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics 81*, 3 (Dec. 2009), 719–745.
17. Reacher, M.H., Shah, A., Livermore, D.M. et al. Bacteraemia and antibiotic resistance of its pathogens reported in England and Wales between 1990 and 1998: Trend analysis. *British Medical Journal 320*, 7229 (Jan. 2000), 213–216.
18. Rosvall, M and Bergstrom, C. Mapping change in large networks. *PLoS One 5*, 1 (Jan. 2010).
19. Rosvall, M. and Bergstrom, C. Maps of information flow reveal community structure in complex networks. *Proceedings of the National Academy of Sciences of the United States of America 105*, 4 (Jan. 2008), 1118–1123.
20. Ruzzo, W.L. and Tompa, M. A linear time algorithm for finding all maximal scoring subsequences. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology* (Heidelberg, Germany, Aug. 6–10). AAAI Press, Boston, 1999, 234–241.
21. Salehie, M. and Tahvildari, L. Self-adaptive software: Landscape and research challenges. *ACM Transactions on Autonomous and Adaptive Systems 4*, 2 (May 2009).
22. Thomson Reuters Web of Science; http://thomsonreuters.com/web-of-science/
23. Wikipedia, Computer Science; http://en.wikipedia.org/wiki/Computer_Science
24. Zaki, M.J. Sequences mining in categorical domains: Incorporating constraints. In *Proceedings of the Ninth ACM International Conference on Information and Knowledge Management* (Washington D.C., Nov. 6–11). ACM Press, New York, 2000, 422–429.

Apirak Hoonlor (apirak.hoo@mahidol.ac.th) is an instructor in the Faculty of Information and Communication Technology at Mahidol University, Bangkok, Thailand.

Boleslaw K. Szymanski (szymab@rpi.edu) is the Claire and Roland Schmitt Distinguished Professor of Computer Science and the Director of the Center for Network Science and Technology at Rensselaer Polytechnic Institute, Troy, NY.

Mohammed J. Zaki (zaki@cs.rpi.edu) is a professor in the Computer Science Department at Rensselaer Polytechnic Institute, Troy, NY.