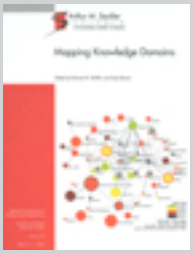


(Sackler NAS Colloquium) Mapping Knowledge Domains

ISBN
978-0-309-09232-6

136 pages
8 1/2 x 11
2004

Proceedings of the National Academy of Sciences

Arthur M. Sackler
COLLOQUIA
OF THE NATIONAL ACADEMY OF SCIENCES

Mapping Knowledge Domains

National Academy of Sciences
Washington, D.C.

Arthur M. Sackler, M.D.
1913–1987

Born in Brooklyn, New York, Arthur M. Sackler was educated in the arts, sciences, and humanities at New York University. These interests remained the focus of his life, as he became widely known as a scientist, art collector, and philanthropist, endowing institutions of learning and culture throughout the world.

He felt that his fundamental role was as a doctor, a vocation he decided upon at the age of four. After completing his internship and service as house physician at Lincoln Hospital in New York City, he became a resident in psychiatry at Creedmoor State Hospital. There, in the 1940s, he started research that resulted in more than 150 papers in neuroendocrinology, psychiatry, and experimental medicine. He considered his scientific research in the metabolic basis of schizophrenia his most significant contribution to science and served as editor of the *Journal of Clinical and Experimental Psychobiology* from 1950 to 1962. In 1960 he started publication of *Medical Tribune*, a weekly medical newspaper that reached over one million readers in 20 countries. He established the Laboratories for Therapeutic Research in 1938, a facility in New York for basic research that he directed until 1983.

As a generous benefactor to the causes of medicine and basic science, Arthur Sackler built and contributed to a wide range of scientific institutions: the Sackler School of Medicine established in 1972 at Tel Aviv University, Tel Aviv, Israel; the Sackler Institute of Graduate Biomedical Science at New York University, founded in 1980; the Arthur M. Sackler Science Center dedicated in 1985 at Clark University, Worcester, Massachusetts; and the Sackler School of Graduate Biomedical Sciences, established in 1980, and the Arthur M. Sackler Center for Health Communications, established in 1986, both at Tufts University, Boston, Massachusetts.

His pre-eminence in the art world is already legendary. According to his wife Jillian, one of his favorite relaxations was to visit museums and art galleries and pick out great pieces others had overlooked. His interest in art is reflected in his philanthropy; he endowed galleries at the Metropolitan Museum of Art and Princeton University, a museum at Harvard University, and the Arthur M. Sackler Gallery of Asian Art in Washington, DC. True to his oft-stated determination to create bridges between peoples, he offered to build a teaching museum in China, which Jillian made possible after his death, and in 1993 opened the Arthur M. Sackler Museum of Art and Archaeology at Peking University in Beijing.

In a world that often sees science and art as two separate cultures, Arthur Sackler saw them as inextricably related. In a speech given at the State University of New York at Stony Brook, *Some reflections on the arts, sciences and humanities,* a year before his death, he observed: "Communication is, for me, the *primum movens* of all culture. In the arts. . . I find the emotional component most moving. In science, it is the intellectual content. Both are deeply interlinked in the humanities." The Arthur M. Sackler Colloquia at the National Academy of Sciences pay tribute to this faith in communication as the prime mover of knowledge and culture.

PNAS
Proceedings of the National Academy of Sciences of the United States of America www.pnas.or

Contents

Papers from the Arthur M. Sackler Colloquium of the National Academy of Sciences

Introduction

Mapping knowledge domains

Richard M. Shiffrin*† and Katy Börner‡

*Psychology Department and ‡School of Library and Information Science, Indiana University, Bloomington, IN 47405

The term "mapping knowledge domains" was chosen to describe a newly evolving interdisciplinary area of science aimed at the process of charting, mining, analyzing, sorting, enabling navigation of, and displaying knowledge. This field is aimed at easing information access, making evident the structure of knowledge, and allowing seekers of knowledge to succeed in their endeavors. Although thousands of years old, this area has undergone a sea change in the last 15 years, a change fostered by an explosion of the amount of information available, the accessibility of that information due to electronic storage, and the new techniques of analysis, retrieval, and visualization that are made possible by vast increases in computational storage capacity and processing speed and power. Many of us are so involved in the new ways of accessing knowledge that we have forgotten how recent is the change to computerized knowledge retrieval with search engines operating on the World Wide Web. Remarkable as these changes are to date, they are only a hint of the transformation to come. The Arthur M. Sackler Colloquium on Mapping Knowledge Domains, held at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA, May 9–11, 2003, was designed to showcase the ongoing developments in this transformation and provide pointers toward the directions it will move.

The changes that are taking place profoundly affect the way we access and use information. Scientists, academics, and librarians have historically worked hard to codify, classify, and organize knowledge, thereby making it useful and accessible. The day is fast approaching when all this knowledge will be coded electronically, but mixed in a vast and largely disorganized and often unreliable sea of mostly recent information. Fishing this sea for desired information is presently no easy task and will continue to increase in difficulty. However, the speed and power of modern computation gives hope that this daunting task can be accomplished. In addition, and perhaps even more important, the new analysis techniques that are being developed to process extremely large databases give promise of revealing implicit knowledge that is presently known only to domain experts, and then only partially.

Some of these techniques are now being applied in science, aiming to identify and organize research areas according to experts, institutions, grants, publications, journals, citations, text, and figures; discover interconnections among these; establish the import of research; reveal the export of research among fields; examine dynamic changes such as speed of growth and diversification; highlight economic factors in information production and dissemination; find and map scientific and social networks; and identify the impact of strategic and applied research funding by government and other agencies. The new techniques support and complement human judgment. They dramatically speed up achievements formerly reached solely by human effort and provide new results that could not have been reached by humans unaided. As the flood of new and disorganized information continues to crest, the new tools are increasingly critical for the growth of scientific research, and indeed for the functioning of modern society.

The importance and fundamental nature of these new ways of interacting with information, and accessing knowledge, have led to considerable interest in for-profit applications. As a result, many of the algorithms and software developed in this field are proprietary. Users are given the end products, such as a list of potentially useful websites or a visual map, without much knowledge concerning the conceptual basis and technical implementation of the underlying algorithms. The desire to promote a deeper understanding therefore led us to include leading researchers not only from academia and government, but also from businesses such as Google and Microsoft.

We thought it would prove useful and interesting if some of the techniques used to map knowledge were applied to the contents of PNAS itself. Thus, we arranged for registered participants to have access to an electronic compilation of the full text documents from PNAS covering January 7, 1997, to September 17, 2002 (148 issues containing some 93,000 journal pages). The time between the first availability of this data set and the deadline for submissions was rather short; nonetheless, several of the contributors analyzed this set, with results that provide interesting directions for future research.

The value of mapping knowledge domains of course extends well beyond the bounds of information science or the PNAS journal, to scientists, researchers, governmental institutions, industry, and members of society generally. It should be emphasized that, although the extraction and organization of knowledge may form the scientific core of this field, the results will be of little use unless the user can understand and interact with the mapping systems. Knowledge typically is organized along many thousands of dimensions, but a map with thousands of dimensions cannot be used effectively by humans. For this reason, domain visualizations and the ability to interact with knowledge and view it from a variety of perspectives play a critical role. The results of algorithms used to extract and organize relevant data can be displayed in many complementary ways. For example, maps might depict major researchers, most cited articles and books, articles too new to receive many citations but with contents that point to emerging trends, articles organized into topic trees (by content, citations, and authors), and grants awarded by topic. Other maps might depict changes over time. Such techniques hold out the promise that the user will be able not only to visualize a few nearby trees in the forest of knowledge, but also to understand the entire landscape. If these techniques can be made to operate effectively, they may well change the way that science is conducted and the way the business of the world is carried out.

Achieving such results requires tools from diverse areas of science: ways to analyze truly enormous amounts of data and extract meaningful results; ways to sort and cluster information

This paper serves as an introduction to the following papers, which result from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

†To whom correspondence should be addressed. E-mail: shiffrin@indiana.edu.

by similarity and importance; ways to identify close and distant interconnections that are not immediately obvious (especially when terminology differs); ways to display large amounts of information that lie along multiple dimensions so that the user can properly interpret the results and guide further exploration; ways to design interactive interfaces; and ways to analyze the structure of the database itself. Examples of many of these are represented in the articles of this special issue, and additional examples were presented at the colloquium (slides and audio files of the talks and a video of the poster presentations are accessible at http://vw.indiana.edu/sackler03/). In the long run, the promise of this field will not be realized without dynamically interactive systems; several of our participants have been developing such systems, but the pages of a journal do not, unfortunately, afford access to dynamic systems.

Overview of Contributions

This special PNAS issue contains three articles that set the stage by providing general coverage of methods, techniques, and practices: an analysis of knowledge extraction from the World Wide Web by Monika Henzinger and Steve Lawrence (Google research); an analysis, correlation, and mapping of paper and grant data to assess research by Kevin W. Boyack (Sandia National Laboratories, Albuquerque, NM); and an analysis of scientific collaboration networks by Mark Newman (University of Michigan, Ann Arbor).

Several articles address methods to extract and organize information from large unstructured databases: Simon Dennis presents an unsupervised method to extract propositional information from a "tennis article" database and answer questions about information implicit in the data. Three articles present methods to organize databases in terms of the semantic similarity of the contents and to apply the methods to the PNAS database. Tom Landauer, Darrell Laham, and Marcia Derr use "Latent Semantic Analysis." Elena Erosheva, Stephen Fienberg, and John Lafferty use a form of mixed-membership model, and Tom Griffiths and Mark Steyvers present another form they term the "Topics Model" (both are a generalization of "Latent Dirichlet Allocation"). Paul Ginsparg, Paul Houle, Thorsten Joachims, and Jae-Hoon Sul classify research areas inherent in a large text database of physics articles by using a "support vector machine."

Other articles assume a knowledge database represented or representable as a graph structure. Dennis Wilkinson and Bernardo A. Huberman present a stochastic network partitioning procedure that extends the "Girvan–Neuman" method to large, complex graphs to account for nodes that belong to several clusters. They use it to identify communities of genes related to colon cancer. The method could as well be used to determine communities of papers or authors from paper-citation or coauthor networks.

Most data sets evolve over time, and several articles address ways to track dynamic changes in structure. One approach analyzes the changes in users' interactions with the database. The article by John Hopcroft, Omar Khan, Brian Kulis, and Bart Selman applied a new clustering algorithm to the NEC CiteSeer database that would identify real changes in structure without identifying changes due to random and small perturbations. Jonathan Aizen, Daniel Huttenlocher, Jon Kleinberg, and Antal Novak present a stochastic method to analyze the dynamics of item popularity (web traffic) on the Internet Archive and use it to identify points of significant change in real world events.

Given the complexity and nonlinearity of the structure and evolution of, for example, article networks, coauthor networks, and web page graphs, those who wish to mine such networks for knowledge must understand the processes by which such networks evolve. Filippo Menczer introduces a mixture model that grows web page or paper-citation networks on the basis both of based existing interlinkages (popularity) and of the content of the papers and web pages, thereby reproducing network "degree" and content similarity distributions. Katy Börner, Jeegar Maru, and Robert Goldstone present a simple process model that simultaneously grows coauthor and paper-citation networks; the statistical and dynamic properties of the simulated network data are validated against a 20-year PNAS data set.

Methods to display and visually explore the results of large-scale database analyses (i.e., visualization techniques) are of critical importance, and these can benefit from centuries of work in geographic metaphors and cartographic techniques. André Skupin reviews major cartographic principles and by way of example produces a large-format map-like knowledge-domain visualization. Alan MacEachren, Mark Gahegan, and William Pike combine geographic visualization techniques and concept mapping to design a tool that helps individual researchers describe the process of knowledge construction and enables teams of collaborators to synthesize common concepts. Ketan Mane and Katy Börner identify and correlate highly frequent and highly bursty words in the PNAS database to visualize the structure and evolution of major research topics over time. Steven Morris and Gary Yen introduce Crossmaps, a technique that can be applied to visualize multiple, overlapping relations among documents such as author collaboration groups vs. topics or research fronts covered by those documents. Howard White, Xia Lin, Jan Buzydlowski, and Chaomei Chen present a tool (and apply it to the PNAS database) that automatically and rapidly generates small-scale, "local" pathfinder networks and self-organizing maps as interfaces for document retrieval. The interfaces show co-cited authors or co-occurring subject headings that can be explored interactively. Chaomei Chen presents a technique to interrelate and visually present network structures of, say, paper-citation or coauthor networks, generated for different time slices.

Opportunities and Challenges

The increasing flood of digitally available data demands the development of sophisticated tools of analysis and display, but the tools are limited by the quality of the data. All of the research described in this special issue requires high-quality data that are both accessible and available in a usable and common format. A good deal of "preprocessing" was in most cases required to transform data to reach this state. We hope to see in the near future the development of tools to take information in different and noisy formats and convert it to a common format, or analyze noisy and inconsistent data directly.

Because present analysis requires clean data, it is often necessary to make use of proprietary databases. There is often a cost of access, a fact that contributes to an increasing "information divide." There are of course efforts to move beyond private information (such as Medline, the ACM digital library, or the Physics E-print Archive), but these are currently unconnected.

Such factors are slowing the development of truly global and freely accessible maps of science, or of general knowledge, but we hope and believe that day will arrive. The Sackler Colloquium on Mapping Knowledge Domains and the present articles that derived from that colloquium provide provocative glimpses of that future day.

Many of our colleagues participated in the organization of this colloquium. We thank the members of our organizing committee: Kevin Boyack, Sandia National Laboratories; Chaomei Chen, Drexel University; Susan Dumais, Microsoft Corporation; Jon Kleinberg, Cornell University; Thomas K. Landauer, University of Colorado; and Josh Tenenbaum, Massachusetts Institute of Technology. We greatly appreciate the time and effort of our additional reviewers; these included many of the authors of the present articles, and also Rex G. Cammack, Blaise Cronin, Tom Erickson, Eugene Garfield, Beth Hetzler, Peter Ingwersen,

Grant Lewison, Francis Narin, Sidney Redner, Ben Shneiderman, and Colin Ware. We also thank Kevin Boyack, Jason Baumgartner, and Ketan Mane for processing and managing access to the Colloquium's PNAS database. Special thanks must also go to Eugene Garfield, the "father" of this field, who gave the keynote at the Colloquium and participated actively in the colloquium, adding greatly to its value. The 5-year PNAS full text data set was provided by Diane M. Sullenberger, executive editor, PNAS. The 20-year PNAS citation data set was extracted from *Science Citation Index Expanded* [Institute for Scientific Information, Inc. (ISI), Philadelphia, PA; Copyright ISI].

Extracting knowledge from the World Wide Web

Monika Henzinger* and Steve Lawrence

Google, Inc., 2400 Bayshore Parkway, Mountain View, CA 94043

The World Wide Web provides a unprecedented opportunity to automatically analyze a large sample of interests and activity in the world. We discuss methods for extracting knowledge from the web by randomly sampling and analyzing hosts and pages, and by analyzing the link structure of the web and how links accumulate over time. A variety of interesting and valuable information can be extracted, such as the distribution of web pages over domains, the distribution of interest in different areas, communities related to different topics, the nature of competition in different categories of sites, and the degree of communication between different communities or countries.

The World Wide Web has become an important knowledge and communication resource. As more people use the web for more tasks, it provides an increasingly representative and unprecedented in scale machine-readable sample of interests and activity in the world.

However, the distributed and heterogeneous nature of the web makes large-scale analysis difficult. We provide an overview of recent methods for analyzing and extracting knowledge from the web, along with samples of the knowledge that can be extracted.

Sampling the Web

The sheer size of the web has led to a situation where even simple statistics about it are unknown, for example, its size or the percentage of pages in a certain language. The ability to sample web pages or web servers uniformly at random is very useful for determining statistics. For example, we can use random URLs to estimate the distribution of the length of web pages, the fraction of documents in various Internet domains, or the fraction of documents written in various languages. We can also determine the fraction of web pages indexed by various search engines by testing the engines for the presence of pages chosen uniformly at random.

Random Walk. One approach to sample web pages approximately uniformly at random is based on the idea of a *random walk*, where we take successive steps in random directions. Henzinger *et al.* (1) have performed several such random walks on the web. Their main idea is to perform a random walk so that a page is visited by the walk with probability roughly proportional to its PageRank (2) value, and then to sample the visited pages with probability inversely proportional to their PageRank value. Thus, the probability that a page is sampled is a constant independent of the page.

One definition of the PageRank value of a web page uses a random walk: *The initial page of the walk is chosen uniformly at random from all pages. Assume the random walk is at page* p *at a given time step. With probability* d, *follow an outlink of paper* p, *chosen uniformly at random. With probability 1 −* d, *select a random page out of all pages.* The PageRank of a page p is the fraction of steps that the walk spent at p in the limit, i.e., the PageRank is the stationary distribution of the random walk.

When trying to implement this random walk to generate random web pages, two problems arise: (*i*) The random walk assumes already that we can find a random page on the web, the very problem that we want to solve. (*ii*) Many hosts on the web have a large number of links within the same host and very few leaving them. If such a host is encountered early in the walk, then there is a good chance that most pages are from this host when the walk is stopped, i.e., the walk "never found its way out of the host." The main culprit is that any implementation can only take a finite number of steps, whereas the definition requires an infinite number.

To avoid these problems, Henzinger *et al.* (1) proposed and implemented the following modified random walk: *Given a set of initial pages, choose one page at random to be the start page. Assume the random walk is at page* p *at a given time step. With probability* d, *follow an outlink of page* p, *chosen uniformly at random. With probability 1 −* d, *select a random host out of all hosts visited so far, and jump to a randomly selected page out of all pages visited on this host so far. In this definition, all pages in the initial set are also considered to be visited.*

The two problems are avoided as follows: (*i*) Instead of choosing a random page out of all pages, a random page from a subset of visited pages is chosen. (*ii*) Instead of jumping to a random page, the walk jumps to a random host and then to a random visited page on that host. In this way, even a host that has dominated the walk so far only has the same chance of being visited as any other visited host.

Because of the modification in the walk and because of the fact that the walk has to be finite in practice, the modified random walk visits a page with probability approximately proportional to its PageRank value, which is the stationary distribution of the PageRank random walk.

Afterward, the visited pages are sampled with probability inversely proportional to their PageRank value. If the PageRank value is not known, it can be approximated by computing PageRank on the graph of visited pages. Alternatively, the *visit ratio*, i.e., the ratio of the number of times the page was visited over the length of the random walk, can be used as an approximation of the PageRank value. The latter holds because the PageRank value of a page is defined to be the visit ratio of the PageRank random walk in the limit.

As an example of the statistics we can generate by using this approach, Table 1 shows the percentage of URLS in each top-level domain in fall 1999 generated with this method. Other approaches for sampling web pages based on a random walk methodology are presented in Bar-Yossef *et al.* (3) and Rusmevichientong *et al.* (4).

IP Address Sampling. An approach to obtaining a random sample of web servers is to randomly sample IP addresses, testing for a web server at the standard port (5). There are currently 256^4 ($\approx$4.3 billion) possible IP addresses. If IPv6 was widely used on the web, this approach may not be possible; however, IPv6 has

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

*To whom correspondence should be addressed. E-mail: monika@google.com.

www.pnas.org/cgi/doi/10.1073/pnas.0307528100

Table 1. The top 10 top-level domains according to the percentage of sampled pages in each domain

Domain	Percentage of pages
com	46.93
edu	9.27
org	8.59
net	4.74
jp	3.51
de	3.17
gov	2.92
uk	2.75
ca	1.95
au	1.69
us	1.67
fr	0.81

jp, Japan; de, Germany; uk, United Kingdom; ca, Canada; au, Australia; us, United States; fr, France.

not been widely adopted and this approach is still practical today. Of the 4.3 billion possible IP addresses, some are unavailable and some are known to be unassigned. Many sites are temporarily unavailable due to Internet connectivity problems or web server downtime. To minimize this effect, all IP addresses can be checked multiple times.

This method finds many web servers that would not normally be considered part of the publicly indexable web. These include servers with authorization requirements (including firewalls), servers that respond with a default page, servers with no content (e.g., sites "coming soon"), web hosting firms that present their homepage on many IP addresses, printers, routers, proxies, mail servers, CD-ROM servers, and other hardware that provides a web interface. Many of these can be automatically identified, for example, by using regular expressions.

A number of issues lead to minor biases. The sample corresponds to the subset of servers that are active and respond to requests at the polling times. It is possible for one IP address to host several web sites, multiple IP addresses may serve identical content, and some web servers do not use the standard port. It is common for large sites to use multiple IP addresses that serve the same content (for load balancing and redundancy). This could potentially result in a higher probability of finding larger sites. To minimize the bias, we can use the domain name system to identify multiple IP addresses serving the same content, and consider only the lowest numbered address to be part of the publicly indexable web. Most major sites are not virtually hosted, and few public servers operate on a nonstandard port.

Fig. 1 shows a sample of the results of this approach, showing the distribution of server types found from sampling 3.6 million IP addresses in February 1999 (5). About 83% of servers were commercial, whereas ≈6% of web servers were found to have scientific/educational content (defined here as university, college, and research laboratory servers).

Also analyzed in the same study was metadata usage on the homepages of each server, where the results showed that only 34.2% of servers contained the common "keywords" or "description" metatags on their homepage. The low usage of the simple HTML metadata standard suggests that acceptance and widespread use of more complex standards, such as XML or Dublin Core, may be very slow (0.3% of sites contained metadata using the Dublin Core standard). High diversity was also noted in the HTML META tags found, with 123 distinct tags, suggesting a lack of standardization in usage.

Discussion. Unfortunately, current techniques for sampling web pages exhibit biases and do not achieve a uniform random sample. The main problem with the approaches based on random walks is that any implementation is limited to a finite random walk. The main challenge when using IP address sampling is how to subsample the pages that are accessible from a given IP address.

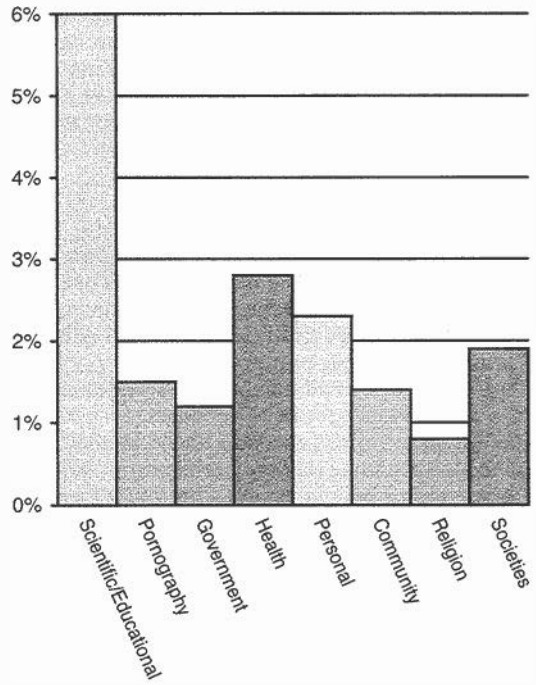


Fig. 1. The distribution of information on publicly indexable web servers. About 83% of servers contained commercial content (e.g., company homepages). The remaining classifications are shown above. Sites may have multiple classifications.

As the web grows it has become impractical to retrieve all pages. Thus, it becomes more important to be able to uniformly sample pages to measure properties of the web. One pragmatic approach is to use two or more approaches that have different biases, for example, a random walk approach and an approach based on IP address sampling, and analyze the agreement between their results.

A fundamental question is what should be counted. For example, consider a web site that contains 10 million pages containing weather statistics for different points in time, compared to another containing the same statistics all on one page. Likewise, a research paper on the web may be on one page or split over multiple pages (6). Additionally, there can be many pages that do not contain original content, they may be transformations of content on other pages (extensions to methods for identify similar document such as ref. 7 can be valuable), or even randomly generated pages. This suggests that some measure of importance may be incorporated into the analysis; for example, we may consider creating a random sample of items that have at least n links to them from other sites, where an item may be a single web page or a collection of web pages (for example, the entire 10 million pages in the weather statistics example). Analysis of web sites as opposed to individual pages is also helpful here.

Analyzing and Modeling Web Growth

We can also extract valuable information by analyzing and modeling the growth of pages and links on the web. Several researchers (8–11) have observed that the link distribution of web pages follows a power law: the probability that a randomly selected web page has k inlinks is proportional to $k - \gamma$, where

$\gamma = 2.1$. The outlink distribution follows a power law with $\gamma = 2.72$. This observation led to the design of various models for the web graph. We describe two models, namely, the preferential attachment model by Barabási and Albert (8, 9) and the copy model by Kleinberg *et al.* (12). We also describe two extensions of these models to better account for deviations of the model from observations.

Preferential Attachment. Barabási and Albert (8, 9) attribute power law scaling to a "rich get richer" mechanism called preferential attachment. As the network grows, the probability that a given node receives an edge is proportional to that node's current connectivity. Specifically, Barabási and Albert propose the following (undirected) web graph model.

Growth. Starting with a small number m_0 of nodes, at every time step add a new node u with $m \leq m_0$ edges.

Preferential attachment. When choosing the nodes to which the new node connects, we assume that the probability p that a new node will be connected to node u depends on the degree k_u of node u, such that $p = k_u/\Sigma_{\text{node w}}\, k_w$.

An analysis based on mean-field theory shows that the probability for a randomly selected node to have k inlinks in this model is proportional to $k - 3$. More specifically, for a node u created at time step t_u, the expected degree is $m(t/t_u)^{0.5}$. Thus, older pages get rich faster than newer pages, leading to a "rich get richer" mechanism.

This model explains the observed power law inlink distribution. However, the model exponent is 3, whereas the observed exponent is 2.1. Additionally, it is not known that older web pages gain inlinks faster than new pages. Finally, different link distributions are observed among web pages of the same category, which we discuss below.

Competition Varies. The early models fail to account for significant deviations from power law scaling common in almost all studied networks. For example, among web pages of the same category, link distributions can diverge strongly from power law scaling, exhibiting a roughly log-normal distribution. In earlier models predicting a power law distribution, most members of a community fare poorly; they have none or very few links to them. However, for actual distributions, many community members can have a substantial number of inlinks, with the mode of the distribution varying up to ≈800 links for universities. Moreover, conclusions about the attack and failure tolerance of the Internet based on the early models may not fully hold within specific communities.

The distributions for outbound web links, and for a variety of other social and biological networks, also display significant deviations from power law (8, 10, 11, 13, 14).

Pennock *et al.* (15) introduced a new model of network growth, mixing uniform and preferential attachment, that accurately accounts for the true connectivity distributions found in web categories, the web as a whole, and other social and biological networks. Previous models imply a drastic "winners take all" scenario on the web, whereby highly referenced pages continue to grow richer in links, whereas new entrants languish in comparison. In fact, the situation is not so inequitable when examined at a local rather than a global level.

Pennock's model generalizes the Barabási–Albert model to incorporate both preferential attachment and a uniform baseline probability of attachment. The model predicts the observed shape of both the body and tail of typical connectivity distributions, including those observed within specific categories of web pages where the divergence from power law is especially marked. In the model, larger modes arise from faster rates of growth of edges as compared to vertices, suggesting an explanation for the different modes observed within different categories of web pages.

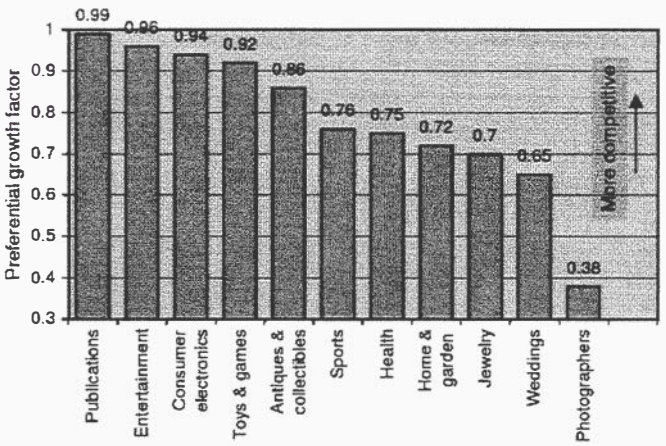


Fig. 2. Competition in different e-commerce categories in March 2002. "More competitive" refers to tougher competition, i.e., it is harder to compete with existing popular sites.

Pennock's model can be used to analyze competition in different categories on the web. Fig. 2 shows the degree of preferential growth for web sites in different e-commerce categories. The publications e-commerce category is the most competitive, where in this case we use competitive to mean that it is harder for a new site to compete with existing sites. The photographers category is the least competitive. There are multiple factors that can lead to the differences in competition that we see. For photographers, one likely factor is their local nature: photographers typically serve only a local community, and those serving different areas usually do not compete. Another factor may be that people looking for photographers use methods other than the web more often (e.g., referrals from friends). Perhaps because people typically use professional photographers rarely, they are also less likely to create and share information among related sites on the web.

A number of models related to Pennock's model have been proposed: Dorogovtsev *et al.* (16) and Levene *et al.* (17) independently propose similar generalizations of the Barabási–Albert model (the addition of a uniform component), motivating it in part as a natural way to parameterize the power-law exponent. Albert and Barabási (18) have proposed their own augmented model that involves a parameterized mixture of three processes: vertex additions, edge additions, and edge rewirings. The combination leads to a connectivity growth function that is roughly a sum of uniform and preferential terms. Even Simon (19) in 1955 invoked a similar process to explain Estoup–Zipf word frequency distributions.

Kleinberg *et al.* (12) explained the power-law inlink distributions with a *copy model* that constructs a directed graph. A slightly modified version as in ref. 20 works as follows: At each time step, one new node u is added with d outlinks. The destinations of these d links with source u are chosen as follows: First, an existing node v is chosen uniformly at random. Then, for $j = 1, 2, \ldots, d$, the jth link of u points to a random existing node with probability α, and to the destination of v's jth link with probability $1 - \alpha$.

Similarly to the Pennock *et al.* (15) model, this model is a mixture of uniform and preferential influences on network growth. A detailed analysis in ref. 20 shows that it leads to a power law inlink distribution as well as to a large number of bipartite cliques.

These models can be used to analyze the fault tolerance of the networks. Recently, Park *et al.* (21) analyzed the Internet for susceptibility to faults and attacks by using simulated data from models similar to those above and with actual data. They find

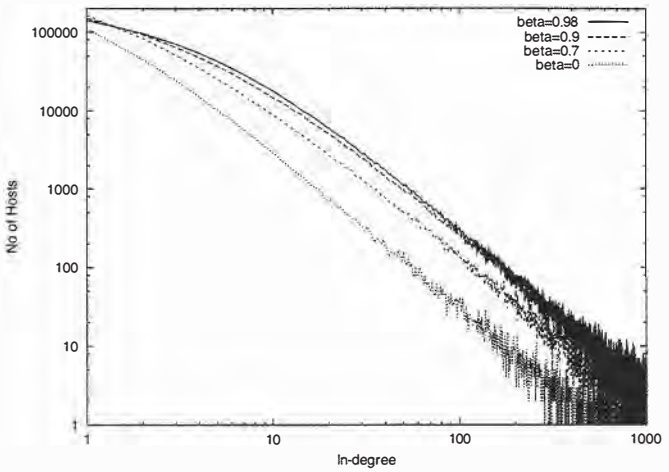


Fig. 3. Inlink distribution as predicted by the "re-link model" with varying β values.

that the Internet is becoming more preferential as it evolves: it is more robust to random failures but is also more vulnerable to attacks.

All of the current models of web growth are an approximation - the true nature of growth on the web is more complex. It is notable that relatively simple models can quite accurately reproduce the actual distributions and behavior of the networks. However, an open problem is refining the models to further improve their accuracy.

The Hostgraph Model. The web is a hierarchically nested graph, with domains, hosts, and pages introducing different levels of affiliation. Instead of modeling the web at the level of pages, one can also model it on the host or domain level. Using the host level leads to the following *hostgraph*: Each node represents a host, and each directed edge represents the hyperlinks from pages on the source host to pages on the target host. Bharat *et al.* (22) show that the weighted inlink and the weighted outlink distributions in the host graph have a power law distribution with $\gamma = 1.62$ and $\gamma = 1.67$, respectively. However, the number of small inlink hosts is considerably smaller than predicted by the model, i.e., there is "flattening" of the curve for low inlink hosts.

Bharat *et al.* (22) present the following modification to the copy graph model, called the *re-link model*, to explain this "flattening": At each time step, with probability β we select a random already existing node u, and with probability $1 - \beta$ we create a new node u. Then we add d new additional outlinks to it. The destinations of these d links with source i are chosen as follows: First, an existing node v is chosen uniformly at random. Second, one picks d random outgoing edges from v. Then, for $j = 1, 2, \ldots, d$, the jth link of u points to a random existing node with probability α, and to the destination of v's jth link with probability $1 - \alpha$.

The difference to the copy model is that with probability $1 - \beta$ no new node is added. Because new nodes start without inlinks the number of low inlink nodes is reduced. Fig. 3 shows the resulting inlink distribution for a graph of 1 million nodes with $d = 7$ and $\alpha = 0.05$ for various β values.

In a recent paper, Cooper and Frieze (23) actually proved that an extension of a model very similar to the re-link model generates graphs whose link distributions follow a power law. Chakrabarti *et al.* (24) used a variant of the Bar-Yossef *et al.* random walk together with a topic classifier to analyze the link distributions of pages on the same topic.

Bharat *et al.* also analyzed affinity between top level country domains in June 2001. Table 2 shows the 20 source domains with the most outlinks together with the .com domain. For each source domain, it lists the percentage of outlinks into the same domain, into the .com domain, and into the four most highly linked country domains from that source domain.

Table 2. Most frequently linked-to domains from country domains

	% of weighted outdegree					
	com	self	1	2	3	4
com	82.9		net 6.5	org 2.6	jp 0.8	uk 0.7
au	27.0	58.8	uk 1.0	ch 0.5	ca 0.4	de 0.3
br	17.8	69.1	uk 0.4	pt 0.4	de 0.4	ar 0.2
ca	19.4	65.2	uk 0.6	fr 0.4	se 0.3	de 0.3
cn	15.8	74.1	tw 0.4	jp 0.2	de 0.2	hk 0.1
cz	8.1	82.4	sk 1.0	de 0.7	uk 0.4	ch 0.1
de	16.0	71.2	uk 0.8	ch 0.6	at 0.5	nl 0.2
dk	13.8	73.0	uk 1.1	de 1.0	int 0.7	no 0.7
es	38.9	42.3	de 1.3	uk 1.0	fr 0.5	int 0.3
fr	20.9	61.9	ch 0.9	de 0.8	uk 0.7	ca 0.5
it	19.3	64.6	de 1.0	uk 0.7	fr 0.4	ch 0.3
jp	17.4	74.5	to 0.8	cn 0.6	uk 0.2	de 0.1
kr	26.5	57.1	jp 0.6	uk 0.5	de 0.3	to 0.3
nl	21.2	61.7	de 1.3	uk 1.1	be 0.6	to 0.5
no	16.1	65.6	de 1.2	se 0.9	uk 0.7	dk 0.6
pl	4.2	92.2	de 0.2	uk 0.1	ch 0.1	nl 0.1
ru	10.0	84.9	ua 0.4	su 0.2	uk 0.2	de 0.2
se	22.6	60.0	nu 1.6	uk 0.9	de 0.7	to 0.6
tw	22.0	66.0	to 1.3	au 0.6	jp 0.6	ch 0.4
uk	34.2	45.9	de 0.7	ca 0.5	jp 0.3	se 0.3
us	34.4	33.1	ca 0.6	uk 0.5	au 0.2	de 0.2

Domains are listed in boldface. au, Australia; br, Brazil; ca, Canada; cn, China; cz, Czech Republic; de, Germany; dk, Denmark; es, Spain; fr, France; it, Italy; jp, Japan; kr, Korea; nl, The Netherlands; no, Norway; pl, Poland; ru, Russia; se, Sweden; tw, Taiwan; uk, United Kingdom; us, United States.

Communities on the Web

The web allows communities to rapidly form with members spread out around the world. Identification of communities on the web is valuable for several reasons. Practical applications include automatic web portals and focused search engines, content filtering, and complementing text-based searches. Community identification also allows for analysis of the entire web and the objective study of relationships within and between communities.

Flake *et al.* (25–27) define a web community as a collection of web pages such that each member page has more hyperlinks (in either direction) within the community than outside of the community (this definition may be generalized to identify communities with varying sizes and levels of cohesiveness). Community membership is a function of both a web page's outbound hyperlinks as well as all other hyperlinks on the web; therefore, the communities are "natural" in the sense that they are collectively organized by independently authored pages. They show that the web self-organizes such that these link-based communities identify highly related pages (Fig. 4).

Identifying a naturally formed community, according to Flake's definition, is intractable in the general case because the basic task maps into a family of nonparametric–complete graph partitioning problems (28). However, if one assumes the existence of one or more *seed* web sites and exploits systematic regularities of the web graph (8, 30, 31), the problem can be recast into a framework that allows for efficient community identification using a polynomial time algorithm.

This is just one of many link-based approaches proposed for identifying collections of related pages. Kumar *et al.* (11) con-

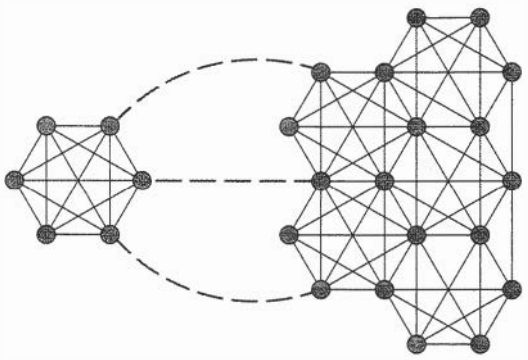


Fig. 4. A community identification example. Maximum flow methods will separate the two subgraphs using any choice of source vertex *s* from the left subgraph and sink vertex *t* from the right subgraph, removing the three dashed links. As formulated with standard flow approaches, all community members must have at least 50% of their links inside the community; however, artificial links can be added to change the threshold from 50% to any other desired threshold. Thus, communities various sizes and with varying levels of cohesiveness can be identified and studied.

sider dense bipartite subgraphs as indications of communities, and expand such graphs into larger sets with HITS (32). Reddy and Kitsuregawa (33) propose a related approach that can be used to identify a hierarchy of communities. Other approaches include bibliometric methods such as cocitation and bibliographic coupling (34–36), the PageRank algorithm (2), the HITS algorithm (32), bipartite subgraph identification (11), spreading activation energy (37), and others (33, 38, 39).

Bipartite subgraph identification, cocitation, and bibliographic coupling are localized approaches that aim to identify well defined graph structures existing in a narrow region of the web graph. PageRank, HITS, and spreading activation energy (SAE) are more global and iteratively propagate weights through a significant portion of the graph. The weights reflect an estimate of page importance (PageRank), how authoritative or hub-like a page is (HITS) or how "close" a candidate page is to a starting region (SAE). PageRank and HITS are related to spectral graph partitioning (40), seeking to find "eigen-web-sites" of the web graph's adjacency matrix or a simple transformation of it. Both HITS and PageRank are relatively insensitive to their choice of parameters, unlike SAE, where results are extremely sensitive to the choice of parameters (37).

Localized approaches are appealing because the structures they identify unambiguously have the properties that the algorithms were designed to find. However, one limitation of these approaches is that they cannot find large related subsets of the web graph because the localized structures are too small. At the other extreme, PageRank and HITS operate on large subsets of the web graph and can identify large collections of web pages that are related or valuable. One limitation of these methods is that it may be hard to understand and defend the inclusion of a given page in the collections that are produced. In practice, HITS and PageRank are combined with textual content either for preprocessing (HITS) or postprocessing (PageRank) (41).

The current approaches to finding communities work well in many, but not all, cases, and have not yet moved from research to widely used products. The approaches often produce some communities with unexpected or missing members. One difficulty is the definition of a community; different people often have different opinions on how a set of pages should be grouped into clusters or communities (29). This is an open area of research.

Summary

The web offers both great opportunity and great challenge in the quest for improving our understanding of the world. The combined efforts of many researchers has resulted in several valuable methods for analysis, and the extraction of a wide variety of valuable knowledge.

However, there are still many open problems and areas for future research. Many of the web analysis studies as presented in this paper provide valuable results for a particular point in time; however, few of these provide directly comparable results at different points in time. It would be very interesting to repeat many of the studies to provide updated analysis, and to provide additional insight into the evolution of the web. The problem of uniformly sampling the web is still open in practice: which pages should be counted, and how can we reduce biases? Web growth models approximate the true nature of how the web grows: how can the current models be refined to improve accuracy, while keeping the models relatively simple and easy to understand and analyze? Finally, community identification remains an open area: how can the accuracy of community identification be improved, and how can communities be best structured or presented to account for differences of opinion in what is considered a community?

1. Henzinger, M., Heydon, A., Mitzenmacher, M. & Najork, M. (2000) *Comput. Networks* **33,** 295–308.
2. Brin, S. & Page, L. (1998) in *Proceedings of the 7th International World Wide Web Conference* (Elsevier, Amsterdam), pp. 107–117.
3. Bar-Yossef, Z., Berg, A., Chien, S., Fakcharoenphol, J. & Weltz, D. (2000) in *Proceedings of the 26th International Conference on Very Large Data Bases (VLDB)* (Morgan Kaufman, San Francisco), pp. 535–544.
4. Rusmevichientong, P., Pennock, D. M., Lawrence, S. & Giles, C. L. (2001) in *American Association for Artificial Intelligence Fall Symposium on Using Uncertainty Within Computation* (Am. Assoc. Artificial Intelligence, Menlo Park, CA), pp. 121–128.
5. Lawrence, S. & Giles, C. L. (1999) *Nature* **400,** 107–109.
6. Eiron, N. & McCurley, K. S. (2003) in *Proceedings of Hypertext 2003* (Assoc. Comput. Machinery Press, New York), pp. 85–94.
7. Broder, A., Glassman, S., Manasse, M. & Zweig, G. (1997) in *Sixth International World Wide Web Conference* (Assoc. Comput. Machinery Press, New York), pp. 391–404.
8. Barabási, A.-L. & Albert, R. (1999) *Science* **286,** 509–512.
9. Barabási, A.-L., Albert, R. & Jeong, H. (1999) *Physica A* **272,** 173–187.
10. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000) *Comput. Networks* **33,** 309–320.
11. Kumar, R., Raghavan, P., Rajagopalan, S. & Tomkins, A. (1999) *Comput. Networks* **31,** 1481–1493.
12. Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S. & Tomkins, A. S. (1999) in *Proceedings of the 5th International Conference on Computing and Combinatorics* (Springer, Berlin), pp. 1–18.
13. Albert, R., Jeong, H. & Barabási, A.-L. (1999) *Nature* **401,** 130–131.
14. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. (2000) *Nature* **407,** 651–654.
15. Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J. & Giles, C. L. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 5207–5211.
16. Dorogovtsev, S. N., Mendes, J. F. F. & Samukhin, A. N. (2000) *Phys. Rev. Lett.* **85,** 4633–4636.
17. Levene, M., Fenner, T., Loizou, Y. & Wheeldon, R. (2002) *Comput. Networks* **29,** 277–287.
18. Albert, R. & Barabási, A.-L. (2000) *Phys. Rev. Lett.* **85,** 5234–5237.
19. Simon, H. A. (1955) *Biometrika* **42,** 425–440.
20. Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A. & Upfal, E. (2000) in *Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE Press, Piscataway, NJ), pp. 57–65.
21. Park, S.-T., Khrabrov, A., Pennock, D. M., Lawrence, S., Gilles, C. L. & Ungar, L. H. (2003) *IEEE Infocom 2003, San Francisco, CA, April 1–3 2003* (IEEE Press, Piscataway, NJ), CD-ROM.
22. Bharat, K., Chang, B.-W., Henzinger, M. & Ruhl, M. (2001) in *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM)* (IEEE Press, Piscataway, NJ), pp. 51–58.

23. Cooper, C. & Frieze, A. (2002) *Random Struct. Algorhythms* **22,** 311–335.
24. Chakrabarti, S., Joshi, M. M., Punera, K. & Pennock, D. (2002) in *Proceedings of the 11th International World Wide Web Conference* (Assoc. Comput. Machinery Press, New York), pp. 517–526.
25. Flake, G. W., Lawrence, S. & Giles, C. L. (2000) in *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining* (Assoc. Comput. Machinery Press, New York), pp. 150–160.
26. Flake, G. W., Lawrence, S., Giles, C. L. & Coetzee, F. (2002) *IEEE Comput.* **35,** 66–71.
27. Flake, G., Tsioutsiouliklis, K. & Tarjan, R. (2002) *Graph Clustering Techniques Based on Minimum Cut Trees* (NEC, New York), Technical Report TR 2002-06.
28. Garey, M. R. & Johnson, D. S. (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, New York).
29. Macskassy, S., Banerjee, A., Davison, B. & Hirsh, H (1998) in *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)* (AAAI Press, New York), pp. 264–268.
30. Huberman, B. A., Pirolli, P. L. T., Pitkow, J. E. & Lukose, R. M. (1998) *Science* **280,** 95–97.
31. Watts, D. & Strogatz, S. (1998) *Nature* **393,** 440–442.
32. Kleinberg, J. M. (1998) in *Proceedings of the Ninth Annual Association for Computing Machinery/Society for Industrial and Applied Mathematics Symposium on Discrete Algorithms* (Assoc. Comput. Machinery/SIAM Press, New York), pp. 668–677.
33. Reddy, P. K. & Kitsuregawa, M. (2002) in *Workshop on Web Analytics, April 13 2002* (Arlington, VA), pp. 11–13.
34. Garfield, E. (1979) *Citation Indexing: Its Theory and Application in Science* (Wiley, New York).
35. Larson, R. (1996) in *Proceedings of the Annual Meeting of the American Society for Information Science* (Assoc. Comput. Machinery Press, New York), pp. 71–78.
36. White, H. D. & McCain, K. W. (1989) *Annu. Rev. Info. Sci. Technol.* **24,** 119–186.
37. Pirolli, P., Pitkow, J. & Rao, R. (1996) in *Proceedings of the Association for Computing Machinery Conference on Human Factors in Computing Systems, Chicago, IL* (Assoc. Comput. Machinery Press, New York), pp. 118–125.
38. Gibson, D., Kleinberg, J. & Raghavan, P. (1998) in *Proceedings of the 9th Association for Computing Machinery Conference on Hypertext and Hypermedia* (Assoc. Comput. Machinery Press, New York), pp. 225–234.
39. Chakrabarti, S., van der Berg, M. & Dom, B. (1999) in *Proceedings of the 8th International World Wide Web Conference* (Elsevier, Amsterdam), pp. 545–562.
40. Chung, F. (1996) *Spectral Graph Theory*, CBMS Lecture Notes (Ar Math. Soc., Providence, RI).
41. Bharat, K. & Henzinger, M. (1998) in *Proceedings of the 21st International ACM SIGR Conference on Research and Development in Information Retrieval* (Assoc. Comput. Machinery Press, New York), pp. 104–111.

Mapping knowledge domains: Characterizing PNAS

Kevin W. Boyack*

Computation, Computers, Information and Mathematics Center, Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185

A review of data mining and analysis techniques that can be used for the mapping of knowledge domains is given. Literature mapping techniques can be based on authors, documents, journals, words, and/or indicators. Most mapping questions are related to research assessment or to the structure and dynamics of disciplines or networks. Several mapping techniques are demonstrated on a data set comprising 20 years of papers published in PNAS. Data from a variety of sources are merged to provide unique indicators of the domain bounded by PNAS. By using funding source information and citation counts, it is shown that, on an aggregate basis, papers funded jointly by the U.S. Public Health Service (which includes the National Institutes of Health) and non-U.S. government sources outperform papers funded by other sources, including by the U.S. Public Health Service alone. Grant data from the National Institute on Aging show that, on average, papers from large grants are cited more than those from small grants, with performance increasing with grant amount. A map of the highest performing papers over the 20-year period was generated by using citation analysis. Changes and trends in the subjects of highest impact within the PNAS domain are described. Interactions between topics over the most recent 5-year period are also detailed.

Scientists have always had the desire to do research of high impact. Part of this desire has been for so-called selfish reasons such as to obtain tenure, increase one's salary, or to enhance one's reputation. However, altruistic purposes also play a large role. We desire to make a difference, to advance knowledge for the benefit of our employers, our nations, or all mankind.

This raises questions that all scientists face and that collectively give rise to innovation and the advancement of science and technology: "What should I work on?" "Are my ideas any good, are they novel, or have they already been taken?" "What can I learn from others?" "How can I improve on their work?" "Who should I work with?" and "Who will fund this?"

Such questions accrue on an institutional level as well. Organizations that answer well are rewarded. Universities develop reputations that drive research agendas and secure large amounts of funding over many years. Successful companies drive markets and consumer preference, maintaining their profitability. Success often reflects an ability to stay on the leading edges of science and technology curves.

In today's world, we have unparalleled access to information, which should enable us to answer questions of a strategic nature more readily than in the past. However, with this increased information has come dilution. Fortunately, tools are now becoming available that allow us to sift, condense, and associate this information in ways that help us answer our questions.

This paper will start with a review of data mining and analysis techniques for the mapping of literatures, including their best uses and the types of questions that can be answered. Subsequent sections will use some of these techniques to provide an indicator-based characterization of the domain comprised by PNAS. Specifically, multiple data sources are combined to give a unique look at input–output (funding–impact) and import–export (diffusion between disciplines) from the perspective of this multidisciplinary, but biomedically dominated journal. A map of the highest impact research in PNAS is also introduced.

Techniques for Mapping Knowledge Domains

Mapping of scientific literature as a field has been in existence for many decades. We are indebted to Eugene Garfield, Derek de Solla Price, and others who, through their desire to understand the structure and flow of scientific advancement (1–5), started the work that has made the indexing and dissemination of bibliographic information a commodity. Electronic sources such as the Science Citation Index Expanded (SCIE), INSPEC, and Medline contain entries for millions of scientific articles, providing us with information to help answer our questions.

Historically, answers have not come without great effort. Given the lack of computing resources, early studies naturally tended to focus on small subsets and were, with some exceptions, academic in nature. With the recent availability of electronic data, exponentially increasing computing power, advanced algorithms, and visualization techniques, we are now at a point where much less effort is required to get answers. Indeed, we can almost routinely do large scale studies aimed at answering significant questions of a strategic nature (6).

Notable among recent advances is the development of the field of information visualization. The past decade has seen rapid growth in this field, and the application of many new techniques to the visualization of literature, patents, genomes (cf. ref. 7), and other information types (8, 9). However, it must be remembered that whereas visualization can be critical to understanding, it is simply a window into the rigorous, often multidimensional, analyses that have formed the basis of informatics for many years. Thus, *mapping*, as a term, does not merely refer to the visualization piece, but to the underlying data mining and analysis techniques as well.

Mapping knowledge domains, then, takes as its input such seemingly diverse subjects as network analysis (e.g., web, social networks, scale-free networks, and metabolic pathways), linguistics, concept or topic extraction, citation analysis, and science and technology indicators, in addition to visualization techniques. Similarly, *knowledge domain* can be more broadly defined than the narrow "technical field" that is commonly associated with the term. Genomes, communities, and networks are all domains with multiple attributes from which one can derive different types of knowledge. Although this paper focuses on mapping of literatures, many of the same analysis and visualization techniques have been and can be applied to other domains.

The main purpose of mapping knowledge domains is to give us knowledge, or answers to our questions. Mapping is useful for

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

Abbreviations: SCIE, Science Citation Index Expanded; ISI, Institute for Scientific Information; ALNR, articles, letters, notes, and reviews; PI, principal investigator.

*E-mail: kboyack@sandia.gov.

www.pnas.org/cgi/doi/10.1073/pnas.0307509100

Table 1. Summary of commonly utilized literature mapping techniques and their uses

Unit of analysis	Questions related to: Fields and paradigms	Communities and networks	Research performance or competitive advantage	Commonly used algorithms
Authors		Social structure, intellectual structure, some dynamics	Use network characteristics as indicators	Social network packages, multidimensional scaling, factor analysis, Pathfinder networks
Documents	Field structure, dynamics, paradigm development		Use field mapping with indicators	Cocitation, co-term, vector space, latent semantic analysis, principle components analysis, various clustering methods
Journals	Science structure, dynamics, classification, diffusion between fields			Cocitation, intercitation
Words		Cognitive structure, dynamics		Vector space, latent semantic analysis, latent dirichlet allocation
Indicators and metrics			Comparisons of fields, institutions, countries, etc., input–output	Counts, correlations

the subject matter expert and nonexpert alike. For the nonexpert, mapping provides an entry point into a domain, a means of gaining knowledge on both the macro and micro levels. For the expert, mapping provides validation of perceptions and a means to quickly investigate trends and new information. Yet, even the expert can be surprised by developments on the periphery of his perception. Mapping and interactive exploration provide context for such surprises.

Commonly utilized techniques for mapping literatures are shown in Table 1 with their primary uses. Most questions of interest fall into three categories: fields and paradigms, communities or networks, and assessment of performance or opportunity. Coauthorship analysis is very similar to social network analysis. Yet, whereas social network analysis is concerned with global properties of large author databases (10), coauthorship studies aim to answer specific questions about collaboration groups (11). Author cocitation analysis is particularly suited to investigation of intellectual structure and history, and is often used with factor analysis and multidimensional scaling (12). Pathfinder network scaling is particularly effective at preparing these data for layout in a visualization program (9).

Documents are the most often used unit of analysis because they can be used to map a particular scientific or technical field and its development. Cocitation and co-word are the two most common types of document analysis, and often lead to different groupings of documents. At the finest levels, cocitation techniques cluster documents by scientific paradigm, or by the same research question and hypotheses (9), whereas co-word document clusters are more topical in nature. Alternatives to the co-word method for generating document similarities include Salton's vector space model (13) and latent semantic analysis (14, 15). Journals are used less often, and are used for larger scale studies, such as to view the relationships between different fields (16). They are also suitable for the study of diffusion between disciplines (often called import–export) by using intercitation rates (17).

Mapping of words or indexing terms as networks reveals the cognitive structure of a field (18). There is some debate as to whether co-word analyses should be used for studies of science dynamics (19). The most reliable approaches aim to combine co-word techniques with citation analyses (20). More advanced techniques using sophisticated algorithms to group and relate topics show great promise for dynamic studies (21, 22).

Similar visualization methods are applied to the mapping types mentioned above for the simple reason that authors, documents, journals, and words (or groupings of these) all work equally well as the mapping unit. Common visualizations include traditional scatterplots and link-node diagrams, such as those drawn by the PAJEK program (23). Newer, more powerful visualizations include self-organized maps (24), landscapes (25, 26), timelines and crossmaps (27), and 3D displays (9). The best of these have the capability of allowing the user to navigate the information space and get detail on demand, which facilitates analysis that helps the user to answer questions.

The power of visualization is enhanced when mapping types are combined. Combining types adds more dimensions to the information, which are more easily explored by using visualization than with traditional analysis methods. For example, Chen (9, 28) combines indicators (citation counts by year) with document cocitation analysis in a 3D display to show the growth of scientific paradigms.

Indicators have been used for as long as people have wanted to compare things. Science and technology indicators were largely developed from the 1950s through the 1970s (29) by the Organization for Economic Cooperation and Development and the National Science Foundation, and have resulted in publications such as National Science Foundation's biannual Science and Engineering Indicators (30). Although activity measures (31), and specifically economic activity measures, have been the dominant component of such reports, scientific output measures such as counts of papers, patents, and citations have also played a large role. Measures of converging partial indicators have been used with the aim of identifying areas of science and technology likely to yield the greatest benefits (32, 33). Output measures have been correlated to economic activity at a macro level to show the relative strengths of countries, states, and/or technical fields (30). Several studies have reported correlation between aggregated scientific outputs and funding (34–39), but none have reported any such correlations at the individual grant level.

Characterization of PNAS

Data Sources. Data from four sources (see Fig. 1) were merged to provide the basis for a characterization of PNAS. Most studies

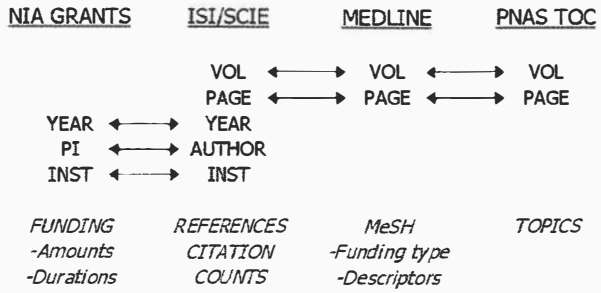


Fig. 1. Data sources, field joins (arrows), and unique properties from each source (italics).

merging databases do so to provide deeper coverage of a field (40, 41). However, this study merges multiple data sources to get more detailed information on a single journal and its impact. The base set to which other sources were merged was data from the SCIE. These data consist of 47,073 records covering the 20 years of PNAS from 1982 to 2001, including full reference lists and citation counts to each paper as of December 31, 2002. Citation counts were determined by matching of Institute for Scientific Information (ISI) reference lists (journal name variations were accounted for) with bibliographic data.[†] For this analysis, only the 45,326 articles, letters, notes, and reviews (commonly referred to as ALNR) were considered. The balance of the records, from editorials, corrections, book reviews, etc., contribute little or no original research, and are commonly discounted in such analyses.

PNAS records were also extracted from Medline, and were joined to the SCIE records primarily for use of the MeSH (medical subject heading) terms. MeSH terms are desirable for several reasons: (*i*) SCIE keywords are sparse, uncontrolled, and available only back to 1991; (*ii*) MeSH is a rich, controlled vocabulary added by human indexers; and (*iii*) MeSH contains specific funding-related terms. Joining MeSH terms to the ISI citation counts enables input–output studies with respect to funding type.

PNAS has a topic structure that is clearly visible in both the print and web versions of the journal Tables of Contents. First-level topics are broad: Biological Sciences, Physical Sciences, and Social Sciences. Within each of these first-level topics are secondary topics, such as Biochemistry, Biophysics, and Cell Biology within the Biological Sciences topic. First- and second-level topics for each paper were extracted from the Tables of Contents and added to the SCIE data. Joining of topics to the other data enables import–export studies as well as the correlation between impact and topic.

Finally, grant data from the National Institute on Aging (one of the institutes of the National Institutes of Health) containing principal investigator (PI) names, institutions, and funding amounts by year were joined to the other data. These data were obtained from the National Institute on Aging as part of a previous study (39). An effort was made to match grants to PNAS papers that were likely to have resulted from specific grants. For a paper to be linked to a specific grant the following conditions were required (also see Fig. 1):

PNAS author = Grant PI (last name + first initial)
and PNAS author institution = Grant PI institution
and PNAS publication year ≥ Grant initial year
and (PNAS publication year ≤ Grant initial year + 5
or PNAS publication year ≤ Grant final year + 2)

[†]These data are extracted from *Science Citation Index Expanded* [Institute for Scientific Information, Inc. (ISI), Philadelphia, PA; Copyright ISI]. All rights reserved. No portion of these data may be reproduced or transmitted in any form or by any means without the prior written permission of ISI.

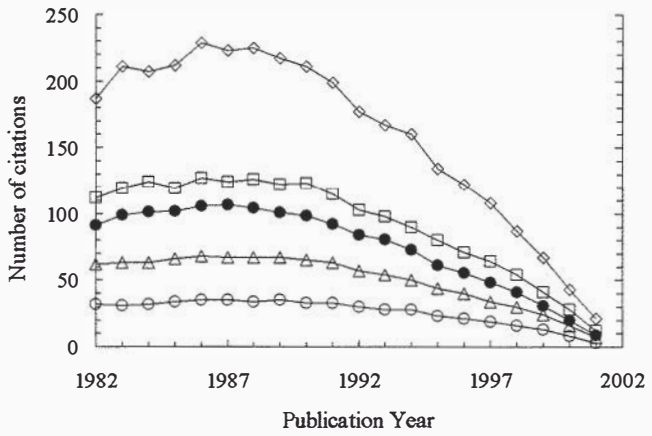


Fig. 2. Mean number of citations (●) to PNAS ALNR are compared with several different percentiles: 90th (◇), 75th (□), 50th or median (△), and 25th (○). Citation counts are as of December 31, 2002.

A total of 1,862 PNAS papers were found to be probable matches to specific grants. Although we cannot say with certainty that these papers are from National Institute on Aging-funded studies, they were authored by National Institute on Aging-funded PIs and were written at a time consistent with their National Institute on Aging funding. Joining of grant data to the balance of the data enables correlation of impact to funding amount, something that has to date been very difficult to quantify.

In this study, *impact* is equated with a ranking measure derived from citation counts. Papers were ranked by citation count for each publication year. Absolute rankings were then converted to percentile rankings. Percentile rankings are used for two reasons. First, it provides normalization across time such that papers from different years can be directly compared. This result is particularly important for recent papers, because they have typically not had enough time after publication to accumulate large numbers of citations. Second, given the skewed nature of citation count distributions, it keeps a few highly cited papers from dominating citation statistics. For example, mean citation counts for the PNAS papers range between the 64th and 70th percentile from 1982 to 1999. Related data are shown in Fig. 2.

Whereas there are certainly factors other than citation measures in what constitutes a full definition of *impact*, and while the validity of using citation measures has been debated (cf. refs. 42 and 43), they are widely used (44), and will be the basis for impact in this study.

Impact and Funding. Medline MeSH terms contain three main funding source designators: *Support, U.S. Gov't, P.H.S.*, *Support, U.S. Gov't, Non-P.H.S.*, and *Support, Non-U.S. Gov't*. The first two designators refer to publications funded by the U.S. Public Health System (P.H.S.) and all other U.S. government agencies (OG), respectively. In a practical sense, P.H.S. refers to the National Institutes of Health. *Support, Non-U.S. Gov't* (nG) could refer to either U.S. nongovernmental sources (e.g., industry, nonprofit) or to foreign sources, but has not been segmented further. Papers with no funding source designators are tagged as *Unknown*. Very few papers in this category exclude a funding acknowledgment inadvertently (45). Thus, *Unknown* can be considered as a distinct category.

Given that each paper is tagged with anywhere from none to all three of the funding source designators, eight unique funding categories can be constructed. Two of the smaller categories, PHS+OG and PHS+OG+nG, have been combined to make a category of sufficient size for statistical purposes. Thus, seven

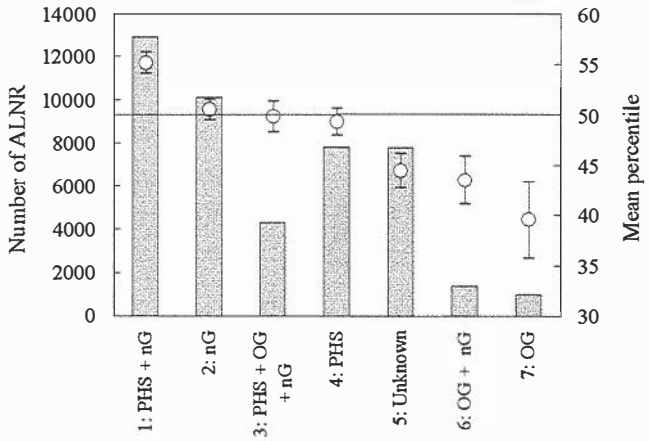


Fig. 3. Numbers of papers (ALNR) and impact (mean citation percentile) for seven funding categories. Categories are shown in order of decreasing mean percentile. Bars indicate the number of papers (*Left*); circles and standard error bars indicate impact (*Right*). PHS, U.S. Public Health System; OG, other U.S. government; nG, non-U.S. government (includes foreign).

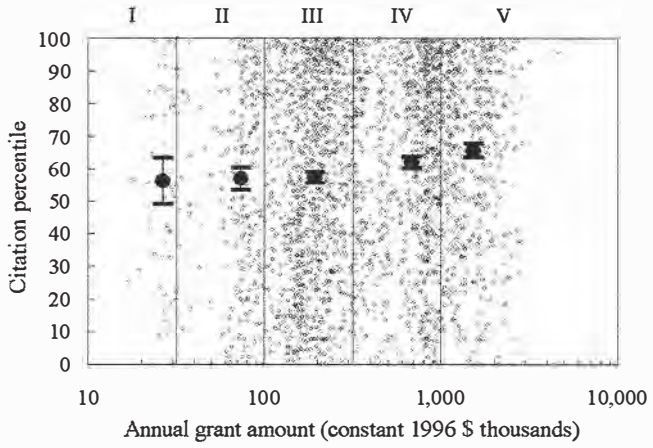


Fig. 4. Correlation between impact (citation percentile) and grant amount. Individual grant-paper pairs (small circles) and mean percentiles with standard errors (large circles) are shown for the five grant size regions that are numbered I–V.

funding categories are shown in Fig. 3 along with their numbers of papers (ALNR) and mean percentiles. The highest ranked category, with a mean percentile >55, is papers jointly funded by the U.S. Public Health System and non-U.S. government sources. By contrast, papers funded solely by the U.S. Public Health System have a mean percentile of 49.2. Yet, this is still higher than the mean percentile of 44.4 associated with papers of *Unknown* funding source, indicating that PHS funding has a positive impact with respect to a lack of U.S. Public Health System funding. The differences between impacts of these three categories are statistically significant at the $P < 0.001$ level by using a Scheffé test (ref. 46 and Table 2).

Other studies have shown that the mean impact of a group of papers increases with the number of authors, presumably due to multidisciplinarity (36). In general, the number of authors increases with the increasing percentile in Fig. 3. However, there are local differences that cannot be explained by number of authors. For example, for categories 1 and 2 (4.82 and 5.04 authors, respectively), and categories 4 and 5 (3.99 and 4.11 authors, respectively), the mean number of authors is anti-correlated with mean percentile.

Fig. 3 shows only mean percentiles for the entire 20-year period of study. Mean percentiles by year are relatively stable for the larger funding categories. Smaller categories showed much more scatter by year.

Does Grant Size Matter? As previously mentioned, the correlation between impact and the amount of funding has historically been difficult to quantify. This correlation is largely due to the difficulty of accurately linking funding information with the publications resulting from those funds. Agencies and institutions, although they track many things, are uniformly poor at keeping track of input–output linkages.

Table 2. Scheffé test results for comparisons between percentile means of different funding categories (from Fig. 3)

Category	2	3	4	5	6	7
1	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$
2		NS	NS	$P < .001$	$P < .001$	$P < .001$
3			NS	$P < .001$	$P < .001$	$P < .001$
4				$P < .001$	$P < .001$	$P < .001$
5					NS	$P < .001$
6						$P = .085$

NS, no significant difference between means.

A total of 1,862 PNAS papers were identified as likely having resulted from National Institute on Aging funding. We assume this to be a small fraction of the total number of National Institute on Aging-funded papers, although the exact fraction is not known. Yet, the number deduced here is consistent with the relative sizes of the National Institute on Aging and the National Institutes of Health.[‡] Many of these papers can be matched to multiple grants, and conversely, many of the grants seem to have given rise to multiple papers. For these data, we have identified 3,059 grant-paper pairs. This finding corresponds well to what we know to be true in research; in many cases, institutions receive multiple grants in complementary areas, and certainly the work from a single grant can spawn more than one publication. Multiple linkages between papers and grants indicate a concentration of activity at an institution. The more money received by a particular PI from a focused organization such as the National Institute on Aging, and the more that PI publishes, the more likely it is that the funds and publications are truly linked.

Fig. 4 shows the correlation between citation percentile and average annual grant amount for the 3,059 grant-paper pairs. Dollar amounts were normalized by GDP deflators to remove inflation biases (30). Annual grant amounts were averaged over the publication year of paper and the three previous years. Five different grant amount ranges were identified: <$31,600, $31,600 to $100,000, $100,000 to $316,000, $316,000 to $1,000,000, and >$1,000,000. Mean citation percentiles and grant amounts were calculated for the grant-paper pairs in each of the five grant ranges. The mean citation percentiles remain constant at 56–57 through the first three ranges (up to $316k), then increase to 62 and 65.6 for ranges IV and V.

The number of authors was also considered here as a potentially confounding variable. Cumulative probability density functions of numbers of authors per paper are nearly identical for funding ranges III-V. Thus, number of authors has little impact on the mean percentiles in these funding ranges.

Several observations can be drawn from these data. First, papers from large grants tend to outperform (in terms of mean citation percentiles) those from smaller grants, with the average

[‡]National Institute on Aging funding is ≈6% of the National Institutes of Health total annually. The 1,862 National Institute on Aging papers are 7.4% of the total National Institutes of Health papers in the PNAS data set.

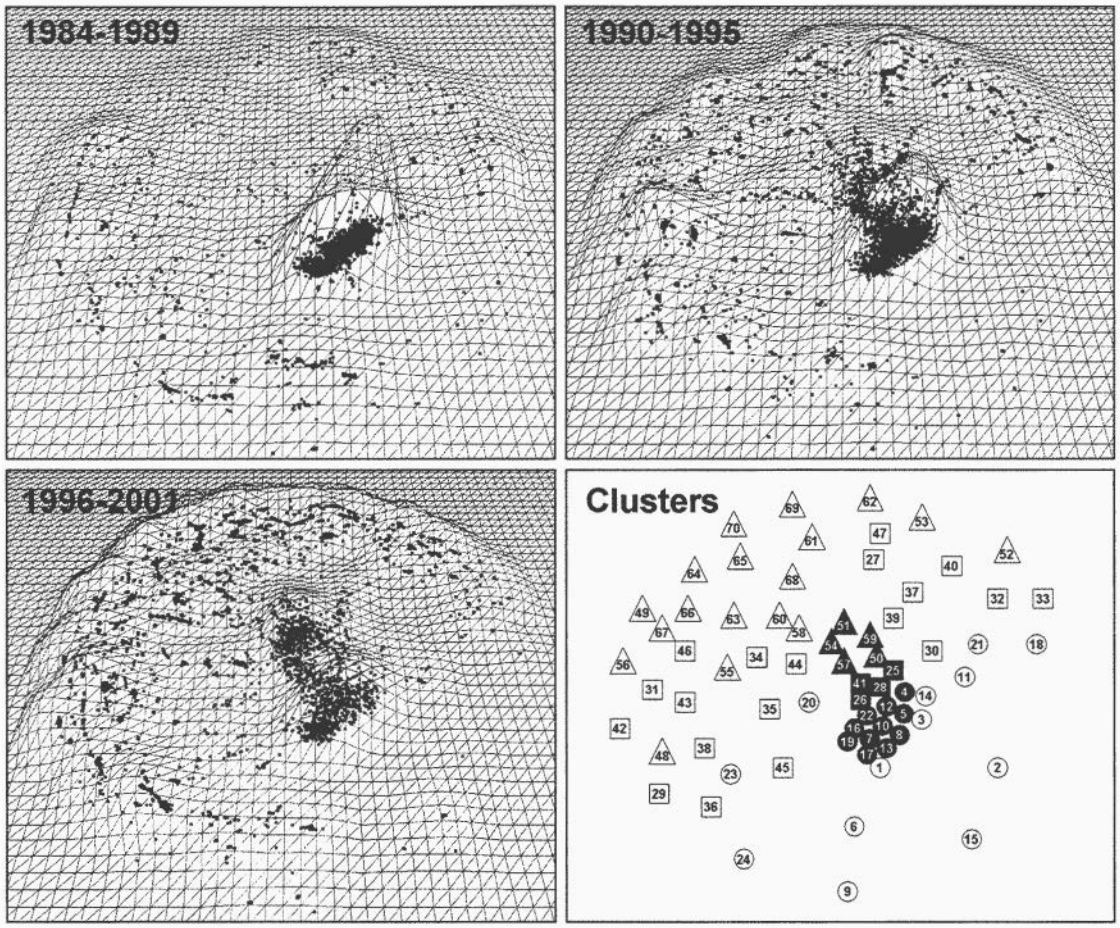


Fig. 5. Three time periods in the PNAS high-impact map show the progression from the basic gene and protein work and techniques that dominated the 1980s to more diverse applications in the 1990s. Maps were generated by using VXINSIGHT. Dots indicate individual papers. Wireframe mountains show the density of papers in clusters. Cluster positions are shown in *Right Lower* for comparison with the map panes. Clusters are numbered from oldest to youngest. Shapes indicate the first third (circles), second third (squares), and last third (triangles) in the timewise progression. Dark shapes indicate the core clusters.

performance increasing with increasing grant amount above $300,000. Second, even for small grants, papers funded by the National Institute on Aging tend to outperform the average PNAS paper; mean percentiles for each grant amount group are well over 50. Third, a high level of funding does not guarantee publication of a high impact paper. Fig. 4 shows many highly funded papers with a low citation percentile. However, the fraction of papers in the lowest quartile for ranges II–V decreases with range (0.199, 0.195, 0.130, and 0.095, respectively), which is consistent with the general increase in mean percentile. Fourth, the variance in individual paper impact appears to be very orthogonal to impact. However, this is to be expected in a single journal study of a high impact journal. If lower impact journals were included in the study, the percentile ranking for most PNAS papers would be shifted much higher.

These observations are specific to National Institute on Aging funding and PNAS papers, and cannot be directly applied to other funding sources or journals. Neither can we claim any direct cause and effect between funding and impact in the results shown here. However, this work shows a similar qualitative correlation between government funding and impact to what has been observed before. Early work by Narin and coworkers (34, 35) showed a positive correlation between National Institutes of Health funding amounts and biomedical publication counts, but did not address impact or quality. Lewison and Dawson (36) used the U.K. Research Outputs Database to show that the mean impact for groups of papers in gastroenterology increased with increases in the number of authors and the number of funding sources. They also found that papers acknowledging funding sources had significantly higher impact than those without such acknowledgments (37). Butler (38) found that whereas acknowledgment data on the whole accurately reflected the total research output of a funding body, there was no ability to track research back to the grant level.

This work goes further than any previous studies by correlating impact with funding level. However, it is also clear that the data are not yet sufficient to produce any definitive conclusions. Government agencies will need to create a clean and maintainable database linking grants, supported publications, patents, and policy changes to enable such analyses (39, 44). Accurate data would enable causal mechanisms to be addressed, given the temporal nature of the grant-research-publication relationship, and would also allow the overall impact (over all publications) of individual grants to be calculated. Such data have the potential to change the way research is funded.

Map of High-Impact Research. To round out this characterization of research published in PNAS, a map was generated to provide information about the subjects of highest impact and related trends. Mapping of all 45,326 ALNR based on their 1.52 million references exceeded the resources available on a common desktop PC. However, a map based on the top quartile of papers from each year, those with a citation percentile of 75 or greater (see Fig. 2), could be easily generated using those same resources.

Table 3. Diagnostic terms and dominant topics for the 50 largest (of 70) clusters from the PNAS high-impact map

Cluster	Mean Year	No. of papers	MeSH term 1	MeSH term 2	Dominant PNAS topic 1997–2001, %
3	1987.40	242	*Oncogenes	DNA restriction enzymes	Biochemistry (30.8)
4	1987.79	483	*Genes, structural	DNA restriction enzymes	
5	1987.82	524	Cloning, molecular	Nucleic acid hybridization	Genetics (37.5)
6	1988.17	281	Oxidation-reduction	Lipoproteins, LDL/*metabolism	Medical Sciences (36.4)
7	1988.46	339	Electrophoresis, polyacrylamide gel	Alzheimer's disease/*pathology	Biochemistry (33.3)
8	1988.80	194	Mutation	Collagen/metabolism	Medical Sciences (33.3)
9	1988.93	94	Buthionine sulfoximine	Bacteriorhodopsins/genetics/*metabolism	Cell Biology (33.3)
10	1988.96	348	Nucleic acid hybridization	Escherichia coli genetics	*Biochemistry (26.7)*
12	1989.25	492	Cloning, molecular	Sequence homology, nucleic acid	Microbiology (31.3)
13	1989.29	254	Transforming growth factors		*Biochemistry (20.7)*
14	1989.37	162	DNA restriction enzymes	H-2 Antigens/*genetics	Medical Sciences (50.0)
16	1990.00	313	Chromatography, affinity	Tumor necrosis factor	*Biochemistry (25.6)*
17	1990.74	93	Sarcoma viruses, avian		Biochemistry (34.8)
18	1990.93	127	Neutralization tests	HIV-1/*immunology	Genetics (72.7)
19	1991.03	171	ADP-ribosylation factors	Hemochromatosis/genetics/* metabolism	Medical Sciences (36.4)
22	1991.26	208	*DNA replication		Biochemistry (34.4)
23	1991.41	144	P-glycoprotein	Drug resistance/*genetics	*Cell Biology (21.4)*
24	1991.45	130	Autoradiography	Receptors, opiod/*metabolism	Biochemistry (32.4)
25	1991.99	193	Chromosome mapping		*Genetics (16.7)*
26	1992.20	172	Receptors, fibroblast growth factor	Receptors, calcitriol	Biochemistry (33.3)
28	1992.44	272	Gene expression	Gene library	Biochemistry (34.5)
29	1993.35	203	Electric conductivity	Synapses/*physiology	Neurobiology (62.5)
31	1993.77	117	*Nucleic acid conformation		*Biochemistry (25.5)*
32	1993.87	157	HIV-I reverse transcriptase	*Reverse transcriptase Inhibitors	Biochemistry (39.5)
36	1994.58	304	Alzheimer's disease/*metabolism	Amyloid β protein/*metabolism	Neurobiology (35.7)
38	1994.78	200	Phosphotyrosine	Protein-tyrosine kinase/*metabolism	*Medical Sciences (22.2)*
40	1995.05	137	Phylogeny	Bone marrow cells	*Evolution (23.5)*
41	1995.10	229	Comparative study	Sequence homology, amino acid	*Medical Sciences (18.0)*
42	1995.10	90	Magnetic resonance imaging	Photic stimulation	Neurobiology (44.9)
43	1995.12	263	Nitric oxide/ *metabolism	ω-*N*-Methylarginine	Medical Sciences (38.7)
46	1995.32	155	Brain-derived neurotrophic factor	Nerve tissue proteins/*pharmacology	Neurobiology (45.8)
47	1995.42	234	*Cell cycle	*Genes, p53	Cell Biology (31.9)
48	1995.54	92	Photosynthetic Reaction Center, bacterial	*Bacterial proteins	Neurobiology (32.6)
49	1995.64	150	*Protein folding	*Protein conformation	Biophysics (69.4)
50	1995.65	302	Molecular sequence data	*Genetic vectors	*Biochemistry (22.8)*
52	1996.09	156	Cytotoxicity, immunologic	Killer cells, natural/ *immunology	Immunology (53.4)
53	1996.10	200	Lymphocyte transformation		Immunology (33.0)
57	1996.54	176	RNA, messenger/genetics/metabolism	Defensins	*Biochemistry (19.3)*
59	1996.86	173	DNA primers	Tetracycline/*pharmacology	*Biochemistry (22.8)*
60	1997.00	82	clF-2 kinase	NF-κ B/*antagonists & inhibitors	Immunology (33.3)
61	1997.21	227	*DNA repair	Leptin	*Medical Sciences (28.8)*
62	1997.35	215	Protein p53/*metabolism	*Genetics, population	*Medical Sciences (24.6)*
63	1997.45	183	Sirolimus	1-phosphatidylinositol 3-Kinase/metabolism	*Cell Biology (27.3)*
64	1997.63	286	*Apoptosis	Protooncogene proteins c-bcl-2	*Cell Biology (24.5)*
65	1997.92	139	Ubiquitins/*metabolism	Multienzyme complexes/*metabolism	*Cell Biology (29.6)*
66	1997.93	205	Models, molecular	Crystallography, x-Ray	Biochemistry (49.0)
67	1997.94	120	Neoplasm transplantation	Serine endopeptidases/*metabolism	*Medical Sciences (25.0)*
68	1998.01	123	Adenomatous polyposis coli protein	Genes, APC	*Medical Sciences (22.4)*
69	1998.31	222	Tumor cells, cultured	*Telomere	Biochemistry (31.7)
70	1999.55	162	Gene expression profiling	Oligonucleotide array sequence analysis	*Genetics (27.1)*

Italics indicate topics with <30% dominance of a cluster.

This approach has the added benefit of focusing only on those topics of highest impact over the years. The resulting map contained 11,565 ALNR. Steps used in creating the map were as follows: (*i*) Paper-to-paper similarities were calculated using bibliographic coupling (47) and direct citations by application of the formula of Small (48), which includes normalization. Cocitation and longitudinal coupling were not considered. 1,744,258 pairs of papers (or 2.61% of the possible pairs) were linked through bibliographic coupling (i.e., having at least one common reference). In addition, the 11,565 ALNR had 411,780 references, of which 24,346 were to other papers within the set. Such direct citations were given a weight of 5. Groups of papers that cite similar sets of references are thus positioned together using this method. (*ii*) Paper positions were calculated from the similarities using VXORD, a force-directed placement ordination routine (49). Ordination does not assign a cluster number to each paper, but rather calculates positions for each paper on an *x*,*y* plane. (*iii*) Papers were assigned to clusters by using the k-means routine in MATLAB based on their *x*,*y* locations from step 2. The number of clusters was arbitrarily set at 70, and whereas 70 is not

Table 4. Summary of properties for PNAS topics, 1997–2001

Topic	No. of ALNR	Mean percentile	Times cited	Independence
Medical Sciences (BS)	1,555	60.0	1,614	0.53
Cell Biology (BS)	1,239	57.5	1,206	0.43
Pharmacology (BS)	189	54.3	126	0.33
Plant Biology (BS)	489	53.3	486	0.69
Genetics (BS)	988	51.9	986	0.47
Microbiology (BS)	499	51.7	514	0.50
Neurobiology (BS)	1,358	51.5	1,098	0.72
Physiology (BS)	341	51.2	209	0.41
Immunology (BS)	865	51.0	730	0.67
Biochemistry (BS)	2,586	49.0	2,521	0.64
Developmental Biology (BS)	372	46.6	266	0.46
Applied Biological Sciences (BS)	95	46.5	67	0.15
Biophysics (BS)	640	46.3	798	0.59
Agricultural Sciences (BS)	44	45.3	39	0.64
Computer Sciences (PS)	10	42.5	5	0.00
Evolution (BS)	527	42.1	470	0.61
Chemistry (PS)	253	41.8	208	0.33
Population Biology (BS)	43	39.4	37	0.19
Psychology (SS)	124	33.9	80	0.56
Ecology (BS)	137	33.7	49	0.80
Applied Physical Sciences (PS)	42	33.3	11	0.36
Engineering (PS)	25	31.2	11	0.27
Geophysics (PS)	26	27.5	4	0.50
Anthropology (SS)	83	25.7	74	0.57
Social Sciences (SS)	11	25.1	4	0.75
Geology (PS)	49	24.6	9	0.44
Statistics (PS)	20	22.5	15	0.20
Physics (PS)	46	22.3	21	0.43
Applied Mathematics (PS)	54	16.4	22	0.50
Astronomy (PS)	14	11.2	3	1.00
Mathematics (PS)	42	7.0	5	1.00
Economic Sciences (SS)	15	4.3	3	0.67

BS, Biological Sciences; PS, Physical Sciences; SS, Social Sciences.

necessarily an optimum number, it is sufficient to show a distribution of topics and trends. Relative cluster positions are shown in Fig. 5. (*iv*) VXINSIGHT (50) was used to interactively navigate and query the PNAS high-impact map. Fig. 5 shows landscapes for three different time periods. When used interactively, tools like VXINSIGHT can show the growth and decay of research fronts in a visual way. (*v*) Diagnostic MeSH terms, i.e., those that differentiate one cluster from another, but that are not necessarily the most common terms, were generated for each cluster, and are given in Table 3. Dominant PNAS topics (from the 1997–2001 Tables of Contents) were also found for each cluster (see Table 3).

The high-impact maps of Fig. 5 show two distinct features: a core group of 20 close-knit clusters in the center, and the remaining clusters that are dispersed and focus on individual topics. The central position of the core clusters indicates their centrality to the focus of PNAS over the 20-year period. This core work had much to do with molecular cloning, hybridization, sequencing, and other key techniques during the first 10 years, shifting into more applied work on growth factors, cancers, and gene expression in the middle years (see Fig. 5 and Table 3 to match diagnostic terms to clusters and times). The most recent work in this core area deals with molecular sequencing, RNA, and cell metabolism.

The dispersed clusters do not have a common focus, but most have strong links (through bibliographic coupling) to the core. In general, the shift has been to more applied topics, often using the revolutionary techniques associated with molecular cloning, hybridization, and sequencing, but maintaining a focus on the application. As a result, clusters of activity have focused on such topics as brain-related research, specific gene and protein activity, protein folding, molecular models, and apoptosis, which was identified as a hot topic from the same data by Griffiths and Steyvers (21).

Another interesting shift is shown by the dominant topics in Table 3. One might assume that papers would tend to cluster within PNAS topics, and that authors would cite heavily to papers of the same topic. Over time, this occurrence has proved to be less and less the case. The number of clusters with less than 30% of their papers belonging to a dominant topic has increased over time. This finding indicates either that coupling between PNAS topics is on the increase or that the perceived boundaries between these topics are becoming more fuzzy.

It is also interesting to consider the characteristics of PNAS topics. Topic assignments are made by authors rather than editors, yet both may wish to see characteristics by topic in that it may influence publishing choices. Second-level topics along with their counts and mean percentile rankings are shown in Table 4. The top 14 topics by percentile are all Biological Sciences topics. Medical Sciences and Cell Biology, although being two of the largest categories, rank highest. The largest category, Biochemistry, has a mean percentile of 49. Physical Sciences and Social Sciences categories all have mean percentiles under 50, which is not surprising for a journal centered in biochemistry.

Mapping of literatures in the ways shown here: i.e., generation of visual maps, clustering, and analysis of the evolution of topics over time, is amenable to discipline level or structural studies as well as to the single journal study given here.

Import–Export Within PNAS. Diffusion of information between scientific disciplines is a relatively new topic of study. The largest of these studies to date looked at 644,000 articles from the 1999 CDROM version of the SCIE (17). Fifteen broad categories of science were defined (e.g., *Basic Life Sciences*, *Biology*, *Physics*, etc.), and the percentage of references from each category to the others was calculated. *Physics* was found to be the most independent, whereas *Biology* was nearly as dependent on *Basic Life Sciences* as upon itself.

Import and export between fields can also be investigated within a single multidisciplinary journal such as PNAS. Here, we look at diffusion between PNAS topics as defined in Table 4. The normalized (number of citations to topic divided by the number of citations to all topics) diagonal of the citation matrix (data not shown) is defined as an index of independence (17, 51), and is given in Table 4. A higher independence value indicates a larger fraction of references given to papers within topic. Independence is thought to correlate with the basic or applied nature of a field, with high independence indicating a basic science (17). A reordering of Table 4 by independence reveals that, in general, the topics order themselves from basic to applied. Plant Biology, Neurobiology, Biophysics, and Biochemistry are all more basic fields than Genetics, Developmental Biology, Cell Biology, or Physiology. For comparison, Rinia *et al.* (17) found that for the entire Science Citation Index for 1999, *Basic Life Sciences* had an independence value of 0.63, whereas the more applied *Biology* had a value of 0.36. However, they also found that Clinical Life Sciences had an independence of 0.67. The PNAS Medical Sciences topic has a value of 0.53, indicating that PNAS Medical Sciences papers may be more enabling (ability to export) than medical sciences papers overall. The full citation matrix shows that Medical Sciences receives >10% of the citations from 11 of the other PNAS topics, including the nonbiological Computer Sciences and Applied Mathematics. The most enabling topic, receiving large fractions of citations from multiple topics, is Biochemistry, which is consistent with the common perception that it forms the core of PNAS publications. Chemistry is anomalous in that it cites heavily to Biochemistry and Biophysics, with an independence of 0.33. The corresponding value from Rinia *et al.* (17) is 0.63. Thus, the PNAS Chemistry topic must be an evolved brand of chemistry that has more to do with application of biology than chemistry at large.

Diffusion between PNAS and other journals could also be examined by using a similar analysis on the citations to and from PNAS.

Conclusions

Impact and funding indicators and citation-based maps have been used to provide a characterization of publication in PNAS from 1982 to 2001. The types of maps and analysis shown here can be applied at many levels: single journal, single discipline, groups of disciplines, etc., given appropriate data. Accurate funding data, and especially, accurate records of the relationship between individual grants and papers is needed. Given these data, similar analyses could be performed for large fields of science, or perhaps, even the whole of science. The ultimate goal is to provide an interactive means of exploring and evaluating scientific and technical information (publications, grants, etc.) to help us obtain answers to questions of strategic importance and aid the innovation process.

I thank the Laboratory Directed Research and Development Program, Sandia National Laboratories, and Katy Börner, Richard Shiffrin, and several anonymous reviewers for insightful comments and suggestions. This work was supported by the U.S. Department of Energy under Contract DE-AC04-94AL85000.

1. Garfield, E. (1955) *Science* **122,** 108–111.
2. Garfield, E. (1970) *Nature* **227,** 669–671.
3. Price, D. J. D. (1963) *Little Science, Big Science* (Columbia Univ. Press, New York).
4. Price, D. J. D. (1965) *Science* **149,** 510–515.
5. Carpenter, M. P. & Narin, F. (1973) *J. Am. Soc. Inf. Sci.* **24,** 425–436.
6. Börner, K., Chen, C. & Boyack, K. W. (2003) *Annu. Rev. Inf. Sci. Technol.* **37,** 179–255.
7. Kim, S. K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., Eizinger, A., Wylie, B. N. & Davidson, G. S. (2001) *Science* **293,** 2087–2092.
8. Card, S., Mackinlay, J. & Shneiderman, B. (1999) *Readings in Information Visualization: Using Vision to Think* (Morgan Kaufmann, San Francisco).
9. Chen, C. (2003) *Mapping Scientific Frontiers: The Quest for Knowledge Visualization* (Springer, London).
10. Newman, M. E. J. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 404–409.
11. Glanzel, W. (2001) *Scientometrics* **51,** 69–115.
12. White, H. D. & McCain, K. W. (1998) *J. Am. Soc. Inf. Sci.* **49,** 327–356.
13. Salton, G., Yang, C. & Wong, A. (1975) *Comm. ACM* **18,** 613–620.
14. Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. & Harshman, R. A. (1990) *J. Am. Soc. Inf. Sci.* **41,** 391–407.
15. Landauer, T. K., Laham, D. & Derr, M. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5214–5219.
16. Bassecoulard, E. & Zitt, M. (1999) *Scientometrics* **44,** 323–345.
17. Rinia, E. J., van Leeuwen, T. N., Bruins, E. E. W., van Vuren, H. G. & van Raan, A. F. J. (2002) *Scientometrics* **54,** 347–362.
18. Callon, M. & Law, J. (1983) *Social Science Information* **22,** 191–235.
19. Leydesdorff, L. (1997) *J. Am. Soc. Inf. Sci.* **48,** 418–427.
20. Noyons, E. C. M., Moed, H. F. & Luwel, M. (1999) *J. Am. Soc. Inf. Sci.* **50,** 115–131.
21. Griffiths, T. L. & Steyvers, M. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5228–5235.
22. Erosheva, E., Fienberg, S. & Lafferty, J. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5220–5227.
23. Batagelj, V. & Mrvar, A. (1998) *Connections* **21,** 47–57.
24. Lin, X. (1997) *J. Am. Soc. Inf. Sci.* **48,** 40–54.
25. Boyack, K. W., Wylie, B. N. & Davidson, G. S. (2002) *J. Am. Soc. Inf. Sci. Technol.* **53,** 764–774.
26. Wise, J. A. (1999) *J. Am. Soc. Inf. Sci.* **50,** 1224–1233.
27. Morris, S. A., Yen, G., Wu, Z. & Asnake, B. (2003) *J. Am. Soc. Inf. Sci. Technol.* **54,** 413–422.
28. Chen, C. & Kuljis, J. (2003) *J. Am. Soc. Inf. Sci. Technol.* **54,** 453–446.
29. Godin, B. (2003) *Res. Policy* **32,** 679–691.
30. National Science Board. (2002) *Science and Engineering Indicators 2002* (National Science Foundation, Arlington, VA).
31. King, J. (1987) *J. Inf. Sci.* **13,** 261–276.
32. Martin, B. R. & Irvine, J. (1983) *Res. Policy* **12,** 61–90.
33. Irvine, J. & Martin, B. R. (1984) *Foresight in Science: Picking the Winners* (Frances Pinter Publications, London).
34. Frame, J. D. & Narin, F. (1976) *Fed. Proc.* **35,** 2529–2532.
35. McAllister, P. R. & Narin, F. (1983) *J. Am. Soc. Inf. Sci.* **34,** 123–131.
36. Lewison, G. & Dawson, G. (1998) *Scientometrics* **41,** 17–27.
37. Lewison, G. (1998) *Gut* **43,** 288–293.
38. Butler, L. (2001) *Res. Eval.* **10,** 59–65.
39. Boyack, K. W. & Börner, K. (2003) *J. Am. Soc. Inf. Sci. Technol.* **54,** 447–461.
40. Ingwersen, P. & Christensen, F. H. (1997) *J. Am. Soc. Inf. Sci.* **48,** 205–217.
41. Hood, W. W. & Wilson, C. S. (2001) *J. Am. Soc. Inf. Sci. Technol.* **52,** 1242–1254.
42. Seglen, P. (1997) *Allergy (Copenhagen)* **52,** 1050–1056.
43. Seglen, P. (1997) *Br. Med. J.* **314,** 498–502.
44. Narin, F. & Hamilton, K. S. (1996) *Scientometrics* **36,** 293–310.
45. Lewison, G., Dawson, G. & Anderson, J. (1995) in *5th International Conference of the International Society for Scientometrics and Informetrics*, eds. Koenig, M. E. D. & Bookstein, A. (Learned Information, Medford, NJ), pp. 255–263.
46. Scheffé, H. (1953) *Biometrika* **40,** 87–104.
47. Kessler, M. M. (1963) *Am. Doc.* **14,** 10–25.
48. Small, H. (1997) *Scientometrics* **38,** 275–293.
49. Davidson, G. S., Wylie, B. N. & Boyack, K. W. (2001) in *7th IEEE Symposium Inform Visualization (InfoVis 2001)*, eds. Andrews, K., Roth, S. & Wong, P. C., (IEEE Computer Society, Los Alamitos, CA), pp. 23–30.
50. Davidson, G. S., Hendrickson, B., Johnson, D. K., Meyers, C. E. & Wylie, B. N. (1998) *J. Intell. Inform. Syst.* **11,** 259–285.
51. Urata, H. (1990) *Scientometrics* **18,** 309–319.

Coauthorship networks and patterns of scientific collaboration

M. E. J. Newman*

Center for the Study of Complex Systems and Department of Physics, University of Michigan, Ann Arbor, MI 48109

By using data from three bibliographic databases in biology, physics, and mathematics, respectively, networks are constructed in which the nodes are scientists, and two scientists are connected if they have coauthored a paper. We use these networks to answer a broad variety of questions about collaboration patterns, such as the numbers of papers authors write, how many people they write them with, what the typical distance between scientists is through the network, and how patterns of collaboration vary between subjects and over time. We also summarize a number of recent results by other authors on coauthorship patterns.

It has long been realized that the coauthorship of articles in learned journals provides a window on patterns of collaboration within the academic community. Coauthorship of a paper can be thought of as documenting a collaboration between two or more authors, and these collaborations form a "coauthorship network," such as that depicted in Fig. 1, in which the network nodes represent authors, and two authors are connected by a line if they have coauthored one or more papers. The structure of such networks turns out to reveal many interesting features of academic communities.

Networks are not new to bibliometrics; the field has a long history of the study of citation networks (1, 2), the networks formed by the citations between papers. These are quite distinct, however, from coauthorship networks; the nodes in a citation network are papers, not authors, and the links between them are citations, not coauthorship. The coauthorship network is as much a network depicting academic society as it is a network depicting the structure of our knowledge. And, perhaps because of this, it has received far less attention than have citation networks. Nonetheless, it has much of value to tell us, as recent work has shown.

During the 1990s (and possibly earlier), a number of authors pointed out the potential utility of coauthorship data and in some cases performed small-scale statistical analyses of such things as frequency of coauthored articles by particular authors or authors at particular institutions (3–7). But it was with the advent of comprehensive online bibliographies that construction of complete or near-complete coauthorship networks for entire fields became a realistic possibility. Starting around 2000, several researchers began the construction of large-scale networks representing research in mathematics (8–10), biology, physics, and computer science (11) and neuroscience (10).

In this paper, we look in detail at three particular networks of scientific collaborations and describe some of the patterns they reveal. The networks are:

(*i*) A network of coauthorships of papers in the Medline bibliographical database from 1995 to 1999, inclusive. Medline is a widely used and compendious database of papers covering biomedical research. Biomedical research accounts for the largest part of civilian scientific research by far, dwarfing research in all other subjects put together in terms of expenditure. Any study that excluded biomedicine could not claim to be representative of science as it is practiced today.

(*ii*) A network of coauthorships of physicists assembled from papers posted on the widely used Physics E-print Archive at Cornell University (formerly at the Los Alamos National Laboratory) between 1995 and 1999. Physics has led the way in moving from journal publication to author self-publication in online preprint databases, with preprint publication largely replacing journal publication in some subfields. Preprint databases provide a useful source of up-to-the-minute publication records, although their coverage is less complete than that of professionally maintained databases like Medline.

(*iii*) A collaboration network of mathematicians compiled from databases maintained by the journal *Mathematical Reviews*. Of the networks yet studied, this is probably the most complete and accurate, covering the period from 1940 to the present without any break.

Networks *i* and *ii* were constructed by the author from bibliographic data supplied by the maintainers of the corresponding databases. Network *iii* was constructed by J. Grossman and P. Ion (8) and graciously supplied by J. Grossman.

A number of other papers in this volume describe bibliometric studies of a database of papers that appeared in PNAS over the period 1997–2002. Although it would be possible to construct a coauthorship network from these data, such a network would be less satisfactory for the study of collaboration patterns than the networks studied here. Most authors publish in more than one journal, so that data on publications in a single journal would give an incomplete picture of their authorship patterns. The databases studied in this paper are more complete, although certainly they do not claim to document every paper.

The outline of this paper is as follows. In *Statistical Properties of Coauthorship Networks*, we describe a variety of results derived from analysis of our networks and highlight some differences among the three subjects studied. In *Additional Results*, we summarize some recent additional results obtained by using the same or additional data, including a number of results due to other authors. In *Conclusion*, we give our conclusions. Many of the results reported here have appeared previously in refs. 11–15, as well as a number of other papers, which are cited as appropriate.

Statistical Properties of Coauthorship Networks

A summary of the basic statistics of the three networks studied here is given in Table 1. The largest of the networks, not surprisingly, is the biomedical network, with >1.5 million authors over a 5-year period. Even the mathematics network, which covers a much longer period (≈60 years), comes nowhere close to this size. Clearly, biology dwarfs other subjects in terms not only of spending but also of manpower. The number of papers shows a similar pattern, although we do not have precise data for

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9-11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

*E-mail: mejn@umich.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0307545100

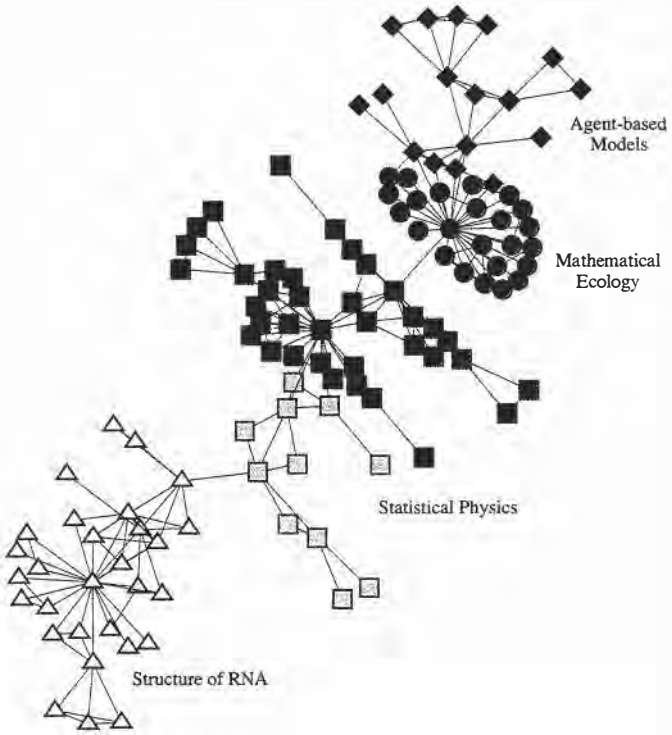


Fig. 1. An example of a small coauthorship network depicting collaborations among scientists at a private research institution. Nodes in the network represent scientists, and a line between two of them indicates they coauthored a paper during the period of study. This particular network appears to divide into a number of subcommunities, as indicated by the shapes of the nodes, and these subcommunities correspond roughly to topics of research, as discussed by Girvan and Newman (37).

Table 1. Summary statistics for the three coauthorship networks analyzed here

	Biology	Physics	Mathematics
Number of authors	1,520,251	52,909	253,339
Number of papers	2,163,923	98,502	—
Papers per author	6.4	5.1	6.9
Authors per paper	3.75	2.53	1.45
Average collaborators	18.1	9.7	3.9
Largest component	92%	85%	82%
Average distance	4.6	5.9	7.6
Largest distance	24	20	27
Clustering coefficient	0.066	0.43	0.15
Assortativity	0.13	0.36	0.12

The statistics are, from top to bottom, total number of authors appearing in the corresponding databases; total number of papers appearing; mean number of papers published by an author; mean number of coauthors on a paper; mean number of different individuals an author collaborated with; largest connected group of individuals in the network; mean vertex–vertex distance between connected individuals in the network; largest such distance; the clustering coefficient, which is the mean probability that two coauthors will also be coauthors of one another; and the degree assortativity coefficient, which is the Pearson correlation coefficient of the degrees (i.e., number of collaborators) of adjacent vertices in the network. The material shown here is after Newman (12) and Grossman (9).

the number of papers in the mathematics database. [Grossman (9) cites a figure of "about 1.6 million authored items" in the *Mathematical Reviews* database, for a slightly more recent version of the network than that studied here.]

The number of papers per author is similar across the three subject areas, between five and seven in each case. Because the mathematics database covers a longer time period, however, this may indicate that mathematicians are producing fewer papers than their more empirically minded colleagues in the sciences. Scientific productivity, measured by number of papers authored, has had a long history of study in bibliometrics, with the articles by Lotka (16) and Shockley (17) being famous early examples. Both of these authors found that the number of papers produced by scientists had a "fat-tailed" distribution, in which a small number of scientists produced a very large number of papers, a result that has since been confirmed by others (18, 19), and which is seen in our own data as well (11, 12).

The number of authors per paper, by contrast, varies substantially among the subjects studied, with biology having the largest number and mathematics the smallest. This presumably reflects real differences in the way research is done in these fields, with biological research consisting often of work by large groups of laboratory scientists and mathematics consisting of theoretical work done primarily by individuals alone or by pairs of collaborators. Grossman (9) says that 66% of mathematics papers are written by a single author (although this number changes over time; see *Additional Results*). In the Medline database, the corresponding figure is 21%. These figures may offer some explanation for the possible lower productivity of mathematics in terms of papers published per unit time: with fewer coauthors on most publications, the production of a mathematics paper involves more work per author.

A similar pattern is revealed in the average number of collaborators an individual has in the three fields, which is more than four times higher in biology than in mathematics. This again is presumably a result of different modes of research, with biology being primarily experimental, mathematics being entirely theoretical, and physics being a combination of the two. [The quintessential example of scientific experiment on an industrial scale is high-energy physics, for which it was previously found, using the SPIRES (www.slac.stanford.edu/spires) database of high-energy physics papers, that authors had an amazing 173 collaborators on average over the 5-year period from 1995 to 1999 (11).]

In addition to mean numbers of papers and coauthors, one can look at the distributions of these quantities. In Fig. 2, for instance, we show the distributions of the number of coauthors that scientists have for the three subjects. The distributions are quite similar, although the distribution for biomedicine (circles)

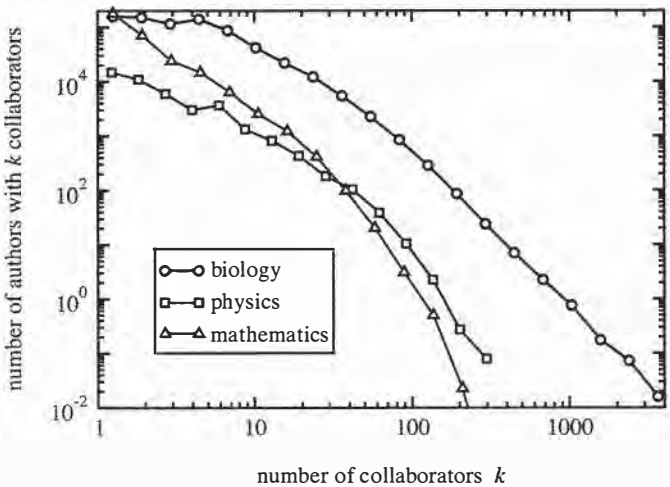


Fig. 2. Histograms of the distribution of numbers of collaborators for scientists in each of three fields studied.

has a longer tail, reflecting the higher mean number of collaborators that individuals have in that field. In each case, the distribution is fat-tailed, like the distribution of number of papers written by scientists mentioned above, with a small fraction of scientists having a very large number of collaborators, up to thousands in the case of the biology network. (Recall that the data for this network cover only a 5-year period; publishing papers with 1,000 coauthors in <2,000 days is an impressive achievement by any measure.) Unlike some other networks, such as the World Wide Web and the Internet, however, the distributions for these networks do not follow power laws; they are not "scale-free networks," in the jargon of the field. It has been suggested that the distributions are actually power law in form with an exponential cutoff (9, 11, 20), and this appears to be a reasonable fit to the data. The cutoff may be produced by the finite time window used in the study, a hypothesis that could, in principle, be tested by varying the size of the window, although we do not do that here.

Table 1 also gives the size of the largest component in each of the networks. A component is a set of network nodes connected via coauthorship, such that any node in the set can be reached from any other by traversing a suitable path of intermediate collaborators. For each of the networks studied here, the largest component fills most of the network, occupying 82–92% of the network in the three cases. Thus a large portion of each of these communities is connected in a kind of linked research enterprise rather than working separately in isolation. Overall, this seems a promising picture; intellectual isolation from the mainstream of one's research area cannot often be a good thing. Most scientists who do not belong to the largest component are members of small disconnected components containing only a handful of others.

Many recent studies of networks of various types have focused on network distance between nodes. This distance is defined as the number of "hops" along links in the network that one needs to make to move from one given node to another. A pair of individuals who have coauthored a paper, for instance, are distance 1 apart, whereas a pair who have not done so but who share a common coauthor are distance 2 apart, and so forth. In the late 1950s, Kochen and Pool (21) speculated on mathematical grounds that networks might show surprisingly small typical distances between pairs of nodes, and in a famous experiment some years later, Milgram (22, 23) demonstrated that this was the case for acquaintance networks, at least in the U.S. Our coauthorship networks appear to follow the same pattern. We calculate the distance between all pairs of individuals in a network using a breadth-first search or "burning" algorithm (24) and then take the average to give the figures shown in Table 1. For each of the networks, the result is very small, at least compared with the size of the network. Mathematics has the largest mean distance, possibly again as a result of the relative sparsity of mathematics collaborations, but even its value of 7.6 is tiny compared with the quarter of a million mathematicians in the network. This appears to indicate a close-knit cohesive community in which most people are connected not only by some path through the network but also by a short one.

A certain amount of attention has focused also on the distances from particular scientists to others in the coauthorship network. Mathematicians have long discussed the "Erdös number," the distance through the mathematics network from a given mathematician to Paul Erdös, an influential Hungarian number theorist of the 20th century who was renowned for his prolific publication and collaboration. Erdös numbers have been studied in depth by a number of authors by using the *Mathematical Reviews* data (8, 9, 25, 26). It is found, for instance, that the mean distance from Paul Erdös to other mathematicians is much lower than the mean distance in the network as a whole, taking a value ≈4.7. [Mean distances from other individuals, most of which are significantly higher than Erdös' mean, are sometimes called "Doe numbers" (9).] The largest distance in a network, which is called the "diameter," is also occasionally of interest; it is between 20 and 30 for each of the networks studied here.

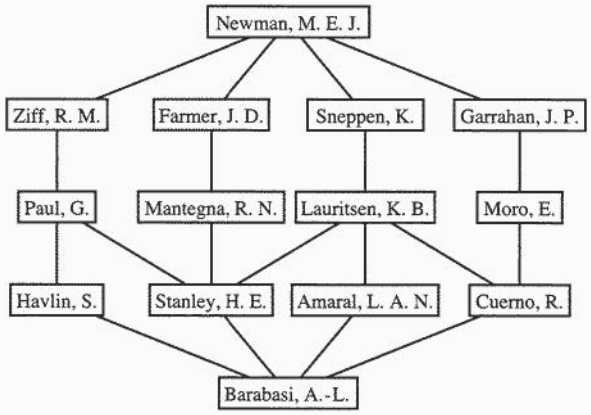


Fig. 3. The shortest paths through the collaboration network of physics papers from the author of this paper to A.-L. Barabási, who also publishes on networks.

It is straightforward to create a computer algorithm to find the shortest path between two particular scientists, again using breadth-first search, and it has been suggested by Kautz *et al.* (27) that such algorithms could be of use for providing "referral chains," links of acquaintances that individuals could use to establish contact with other scientists. In the simplest case, for example, it might be useful to know that one shared a common collaborator with another scientist if one wished to arrange an introduction. Note that the shortest path between two individuals need not be unique, and in fact, it happens quite frequently that there are two or more shortest paths of equal length. Fig. 3 shows shortest paths in the network of the Physics E-print Archive between the present author and A.-L. Barabási of the University of Notre Dame, who also publishes on networks. As Fig. 3 shows, there are several different paths from one scientist to the other, all with length four. This particular case is interesting, because it shows that scientists working in the same field need not be linked through others in their field. The shortest paths in this case are established via my collaborations with J. D. Farmer, J. P. Garrahan, K. Sneppen, and R. M. Ziff, only the last of which collaborations involved work on networks (and then only peripherally).

Another interesting network measure related to shortest paths has been suggested by S. H. Strogatz (personal communication), who asks how many of the shortest paths from a particular individual to others pass through each of their collaborators. Is it the case that most of our connections to others are via just one or two of our best-connected collaborators, or are they distributed evenly among our collaborators? For the networks studied here, it turns out that the former is the case, as is evident in Fig. 4, which shows for the physics network what percentage of shortest paths pass through each of a scientist's coauthors, on average. Thus, Fig. 4 reveals that on average ≈64% of an individual's shortest paths to others pass through the best-connected of their collaborators, and most of the remainder pass through the next-best connected. This may indicate that a small number of scientists are playing the role of broker for communications among others. (See also the discussion of betweenness centrality in *Additional Results*.)

Two other quantities of interest, both previously studied for many networks, are also given in Table 1. The first is the "clustering coefficient" (28), which measures network "clustering" or "transitivity," the probability that two of a scientist's coauthors have themselves coauthored a paper. In topological terms, it is a measure of the density of triangles in a network, a

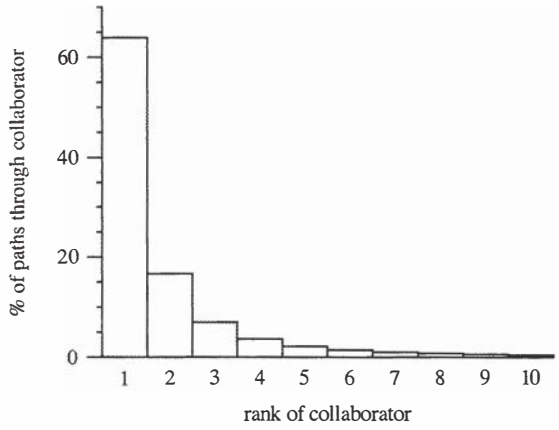


Fig. 4. The average percentage of paths from other scientists to a given scientist that pass through each collaborator of that scientist, ranked in decreasing order. The plot is for the physics network, although similar results are found for the others. [Reproduced with permission from ref. 13 (Copyright 2001, American Physical Society)].

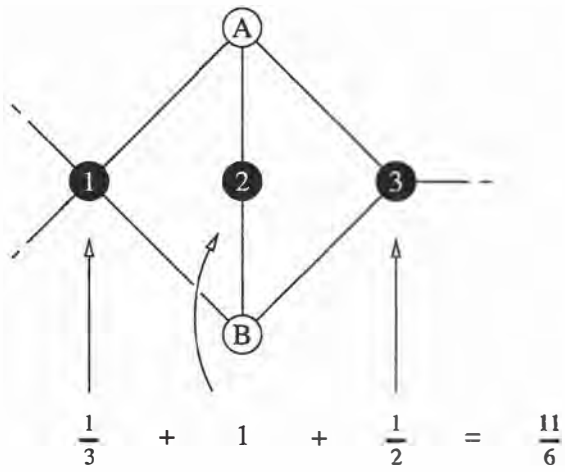


Fig. 5. Authors A and B have coauthored three papers, labeled 1, 2, and 3, which had, respectively, four, two, and three authors. The tie between A and B accordingly accrues weight ⅓, 1, and ½ from the three papers, for a total weight of 11/6. [Reproduced with permission from ref. 13 (Copyright 2001, American Physical Society)].

triangle being formed every time two of one's collaborators collaborate with each other. The clustering coefficient is highest for physics (43%) and lowest for biology (7%), and it is unclear why there is so much variation among fields. Presumably the numbers reflect substantial differences among collaboration patterns in the sciences, but what these differences are is far from obvious. Part of the clustering in each network can be accounted for by papers with three or more coauthors. Such papers introduce triangles of collaborating authors and hence increase the clustering coefficient. This effect can account for only about one-half of the clustering seen in coauthorship networks, however (29); the rest must be due to sociological or organizational effects of some kind.

An alternative way to measure the clustering effect is to look at the time evolution of a network. Among social networks, coauthorship networks are unusual in having well-documented time evolution. Because each paper comes with a date of publication or submission, we can say approximately when each connection was added to the network, and so we can reconstruct the order in which the network grew. This allows us to ask the probability of two scientists coauthoring a paper, given that they have a third mutual collaborator and have not collaborated in the past. By studying only scientists who have not previously collaborated, we eliminate any bias introduced by papers with three or more coauthors. In ref. 14, we showed that scientists with a single mutual collaborator are $\approx$45 times more likely to coauthor a paper than those with no mutual collaborators. Those with two are >100 times as likely to coauthor a paper.

The last line in Table 1 gives the "assortativity coefficient" or degree correlation for the networks (15). This is the correlation coefficient for the number of collaborators that coauthors have. It lies in the range of -1 to 1, with positive values indicating that people with many collaborators tend to collaborate with others who have many collaborators and negative values indicating the reverse. The coefficient is positive for all networks studied, indicating that the most gregarious scientists tend to be connected to each other. Again, it is an open question why this should be the case.

It is also possible to extract from coauthorship data a measure of the strength of the collaboration between pairs of individuals. The simplest such measure would be just a count of the frequency with which two scientists have coauthored papers, the number of coauthored papers over a given interval, for instance. However, this fails to take into account the number of other coauthors on each paper. Presumably two authors who collaborate on a 10-author paper are, in general, working less closely with one another than two who produce a two-author paper with no other help. To account for this effect, we proposed in ref. 13 the measure of collaboration strength illustrated in Fig. 5. Each paper coauthored by a given author pair adds an amount $1/(n-1)$ to the strength of their collaboration, where n is the total number of authors on the paper. The rationale behind this choice is that an author divides his/her time between the $n-1$ other authors with whom he/she works on a paper, and hence the strength of the connection to each of them varies inversely as $n-1$. As an example of this measure, we show in Fig. 6 the coauthors of G. Barkema, one of the author's most frequent collaborators, with line thickness used to indicate the strength of the connections. Clearly, there is considerable variation in connection strength. Over the entire physics network connection, strengths range from a maximum of 34.0 to a minimum of $\approx$0.01.

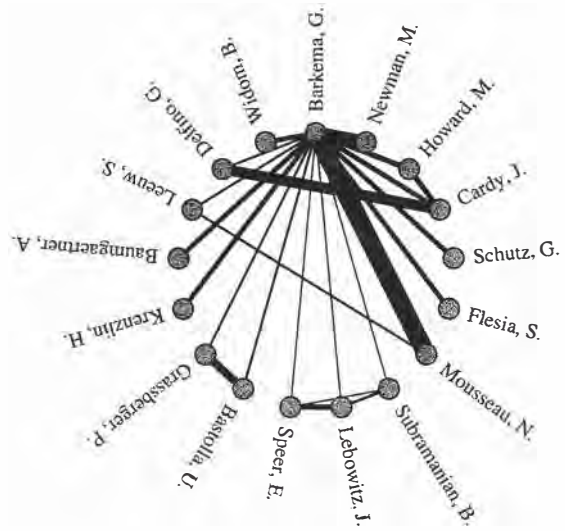


Fig. 6. G. Barkema and collaborators, with lines representing collaborations whose thickness is proportional to the estimate of collaboration strength defined in the text and illustrated in Fig. 5.

Additional Results

The network of collaborations of mathematicians compiled by Grossman and Ion (8) and studied here has also been analyzed extensively by Grossman and collaborators (8, 9, 25) and by others (26). Grossman (9) gives a number of results about the time evolution of the collaboration network. He notes that the rate of publication has increased slightly over the last 50 years or so, but there has been a much more striking increase in the level of collaboration. From the start of the period covered by the *Mathematical Reviews* data in 1940 until the end of the 1950s, less than one-half of all mathematicians had ever coauthored a paper with another writer; nowadays, virtually all of them have. Presumably this trend reflects some combination of changes in the social organization of the mathematics community, better communications, and possibly changes in the types of problems studied and approaches used, making modern mathematics more amenable to collaborative investigation.

The time evolution of coauthorship networks has also been investigated by the present author (14) and others (10) in the context of tests of the "preferential attachment" hypothesis. Price (30) and later Barabási and Albert (31) have suggested that networks grow by the addition of connections in such a way that the probability of an individual gaining a new connection is proportional to the number they already have. Because coauthorship networks are well time-resolved, as discussed in the preceding section, one can test this hypothesis by measuring the probability that a newly published paper contributes new connections to an individual, as a function of the number of connections that individual already has. Refs. 10 and 14 tackle this measurement in slightly different ways, but both conclude that preferential attachment of an approximately linear variety is indeed taking place in collaboration networks.

A number of authors have also looked at "betweenness centrality" in coauthorship networks (13, 32–34). The betweenness centrality of a node A in a network is defined to be the number of shortest paths between other pairs of nodes that pass through A (35). It is regarded as a measure of the influence that individuals have over information flow between others. Individuals who act as brokers for information flow between their colleagues will have high betweenness scores. In ref. 13, it was shown that betweenness scores vary widely from one individual to another in coauthorship networks, with a few having much higher scores than the majority. Later work by Goh *et al.* (33) showed that in fact the distribution of betweenness scores approximately follows a power law. This appears to indicate that collaboration networks contain a small number of influential individuals and many peripheral actors, a conclusion bolstered by the findings of Holme *et al.* (32), who showed that collaboration networks are highly susceptible to the removal of the individuals with highest betweenness scores. One need only remove a few such individuals from the network, it turns out, to break the connection between others and fracture the network into disconnected parts. In a recent paper, Goh *et al.* (34) have extended their investigation of betweenness to the correlation between the betweenness scores of collaborators. They find that there is very little such correlation, implying that influential scientists are not collaborating preferentially with other influential scientists to any significant extent; the probability of one's collaborator having a high betweenness appears not to be significantly greater if one has a high betweenness than if one does not.

Conclusion

In this paper, we have discussed the structure of three networks of scientific collaborations, as deduced from the pattern of coauthorships of papers. The networks cover biomedical research, physics, and mathematics, respectively, and the results reveal both similarities and differences among the different fields. All fields appear to have a broad distribution of the number of coauthors that an individual has, with most individuals having only a few coauthors, whereas a few have many, hundreds or even thousands in some cases. Biological scientists tend to have significantly more coauthors than mathematicians or physicists, a result that reflects the labor-intensive, predominantly experimental direction of current biology. Other differences are less easily explained. In biology, for instance, it is far less likely than in mathematics that two of one's coauthors will also be coauthors of one another, a result that has yet to receive a clear explanation.

Coauthorship networks provide a copious and meticulously documented record of the social and professional networks of scientists. The results reported here represent only a small portion of what could be done with these data. Possible future directions for study might include tests for community structure or "invisible colleges" within the networks (36, 37) or further investigations of changes in collaboration patterns over time (9, 14), as well as other measurements not yet thought of. Coauthorship data represent a superb resource for the pursuit of questions such as these, and I look forward to future developments with interest.

I thank particularly Paul Ginsparg for help in obtaining the data used for this study. The data were generously made available by Oleg Khovayko, David Lipman, and Grigoriy Starchenko (Medline); Paul Ginsparg and Geoffrey West (Physics E-print Archive); and Jerry Grossman (*Mathematical Reviews*). I also thank Steve Strogatz for suggesting the "funneling effect" calculation of *Statistical Properties of Coauthorship Networks* and László Barabási, Paul Ginsparg, Jon Kleinberg, Sidney Redner, Steven Strogatz, and Duncan Watts for useful comments and suggestions. This work was funded in part by the James S. McDonnell Foundation, by the Intel Corporation, and by the U.S. National Science Foundation under Grants DMS-0109086 and DMS-0234188.

1. Price, D. J. (1965) *Science* **149,** 510–515.
2. Egghe, L. & Rousseau, R. (1990) *Introduction to Informetrics* (Elsevier, Amsterdam).
3. Kretschmer, H. (1994) *Scientometrics* **30,** 363–369.
4. Persson, O. & Beckmann, M. (1995) *Scientometrics* **33,** 351–366.
5. Melin, G. & Persson, O. (1996) *Scientometrics* **36,** 363–377.
6. Ding, Y., Foo, S. & Chowdhury, G. (1999) *Int. Inform. Lib. Rev.* **30,** 367–376.
7. Bordens, M. & Gómez, I. (2000) in *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*, eds. Atkins, H. B. & Cronin, B. (Information Today, Medford, NJ).
8. Grossman, J. W. & Ion, P. D. F. (1995) *Congressus Numerantium* **108,** 129–131.
9. Grossman, J. W. (2002) *Congressus Numerantium* **158,** 202–212.
10. Barabási, A.-L., Jeong, H., Ravasz, E., Néda, Z., Schuberts, A. & Vicsek, T. (2002) *Physica A* **311,** 590–614.
11. Newman, M. E. J. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 404–409.
12. Newman, M. E. J. (2001) *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **64,** 016131.
13. Newman, M. E. J. (2001) *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **64,** 016132.
14. Newman, M. E. J. (2001) *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **64,** 025102.
15. Newman, M. E. J. (2002) *Phys. Rev. Lett.* **89,** 208701.
16. Lotka, A. J. (1926) *J. Wash. Acad. Sci.* **16,** 317–323.
17. Shockley, W. (1957) *Proc. IRE* **45,** 279–290.
18. Voos, H. (1974) *J. Am. Soc. Inf. Sci.* (July–August 1974), 270–272.
19. Pao, M. L. (1986) *J. Am. Soc. Inf. Sci.* (January 1986), 26–33.
20. Fenner, T., Levene, M. & Loizou, G. (2002) cond-mat/0209463 (preprint).
21. Pool, I. de S. & Kochen, M. (1978) *Soc. Networks* **1,** 1–48.
22. Milgram, S. (1967) *Psychol. Today* **2,** 60–67.
23. Travers, J. & Milgram, S. (1969) *Sociometry* **32,** 425–443.
24. Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. (2001) *Introduction to Algorithms* (MIT Press, Cambridge, MA), 2nd Ed.
25. de Castro, R. & Grossman, J. W. (1999) *Math. Intell.* **21,** 51–63.
26. Batagelj, V. & Mrvar, A. (2000) *Soc. Networks* **22,** 173–186.
27. Kautz, H., Selman, B. & Shah, M. (1997) *Comm. ACM* **40,** 63–65.
28. Watts, D. J. & Strogatz, S. H. (1998) *Nature* **393,** 440–442.
29. Newman, M. E. J., Strogatz, S. H. & Watts, D. J. (2001) *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **64,** 026118.

30. Price, D. J. (1976) *J. Am. Soc. Inform. Sci.* **27,** 292–306.
31. Barabási, A.-L. & Albert, R. (1999) *Science* **286,** 509–512.
32. Holme, P., Kim, B. J., Yoon, C. N. & Han, S. K. (2002) *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **65,** 056109.
33. Goh, K.-I., Oh, E., Jeong, H., Kahng, B. & Kim, D. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 12583–12588.
34. Goh, K.-I., Oh, E., Kahng, B. & Kim, D. (2003) *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **67,** 017101.
35. Freeman, L. C. (1977) *Sociometry* **40,** 35–41.
36. Crane, D. (1972) *Invisible Colleges: Diffusion of Knowledge in Scientific Communities* (Univ. of Chicago Press, Chicago).
37. Girvan, M. & Newman, M. E. J. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 7821–7826.

An unsupervised method for the extraction of propositional information from text

Simon Dennis*

Institute of Cognitive Science, University of Colorado, Boulder, CO 80301

Recent developments in question-answering systems have demonstrated that approaches based on propositional analysis of source text, in conjunction with formal inference systems, can produce substantive improvements in performance over surface-form approaches. [Voorhees, E. M. (2002) in *Eleventh Text Retrieval Conference*, eds. Voorhees, E. M. & Buckland, L. P., http://trec.nist.gov/pubs/trec11/t11_proceedings.html]. However, such systems are hampered by the need to create broad-coverage knowledge bases by hand, making them difficult to adapt to new domains and potentially fragile if critical information is omitted. To demonstrate how this problem might be addressed, the Syntagmatic Paradigmatic model, a memory-based account of sentence processing, is used to autonomously extract propositional knowledge from unannotated text. The Syntagmatic Paradigmatic model assumes that people store a large number of sentence instances. When trying to interpret a new sentence, similar sentences are retrieved from memory and aligned with the new sentence by using String Edit Theory. The set of alignments can be considered an extensional interpretation of the sentence. Extracting propositional information in this way not only permits the model to answer questions for which the relevant facts are explicitly stated in the text but also allows the model to take advantage of "inference by coincidence," where implicit inference occurs as an emergent property of the mechanism. To illustrate the potential of this approach, the model is tested for its ability to determine the winners of tennis matches as reported on the Association of Tennis Professionals web site.

The closely related fields of question answering and information extraction aim to search large databases of textual material (textbases) to find specific information required by the user (1, 2). As opposed to information retrieval systems, which attempt to identify relevant documents that discuss the topic of the user's information need, information extraction systems return specific information such as names, dates, or amounts that the user requests. Although information retrieval systems (such as Google and Alta Vista) are now in widespread commercial use, information extraction is a much more difficult task and, with some notable exceptions, most current systems are research prototypes. However, the potential significance of reliable information extraction systems is substantial. In military, scientific, and business intelligence gathering, being able to identify specific entities and resources of relevance across documents is crucial. Furthermore, some current information extraction systems now attempt the even more difficult task of providing summaries of relevant information compiled across a document set.

The majority of current information extraction systems are based on surface analysis of text applied to very large textbases. Whereas the dominant approaches in the late 1980s and early 1990s would attempt deep linguistic analysis, proposition extraction, and reasoning, most current systems look for answer patterns within the raw text and apply simple heuristics to extract relevant information (3). Such approaches have been shown to work well when information is represented redundantly in the textbase and when the type of the answer is unambiguously specified by the question and tends to be unique within a given sentence or sentence fragment. Although these conditions often hold for general knowledge questions of the kind found in the Text REtrieval Conference (TREC) Question Answer track, there are many intelligence applications for which they cannot be guaranteed. Often relevant information will be stated only once or may only be inferred and never stated explicitly. Furthermore, the results of the most recent TREC question–answer competition suggest that deep reasoning systems may now have reached a level of sophistication that allows them to surpass the performance possible using surface-based approaches. In the 2002 TREC competition, the POWER ANSWER system (4), which converts both questions and answers into propositional form and uses an inference engine, achieved a confidence weighted score of 0.856, a substantive improvement over the second placed exactanswer (5), which received a score of 0.691 in the main question-answering task.

A key component in the performance of the POWER ANSWER system is its use of the WORDNET lexical database (6). WORDNET provides a catalog of simple relationships among words, such as synonymy, hypernymy, and part-of relations that POWER ANSWER uses to supplement its inference system. Despite the relatively small number of relations considered and the difficulties in achieving good coverage in a hand-coded resource, the additional background knowledge provided by WORDNET significantly improves the performance of the system. This fact suggests that further gains may be achieved if automated methods for extracting a broader range of propositional information could be used in place of, or in conjunction with, the WORDNET database.

In recent years, there have been a number of attempts to build systems capable of extracting propositional information from sentences (7–9).

For instance, given the sentence:

Sampras outguns Agassi in US Open Final,

these systems might produce an annotation such as:

[$_{\text{Winner}}$ Sampras] outguns [$_{\text{Loser}}$ Agassi][$_{\text{Loc}}$ in US Open Final].

This work has been driven, at least in part, by the availability of semantically labeled corpora such as Penn Treebank (10) and FRAMENET (11). As a consequence, the semantic roles used by the systems are those defined by the corpus annotators. However, deciding on a best set of semantic roles has proven extremely difficult. There are a great many schemes that have been proposed ranging in granularity from very broad, such as the two-macro-role

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

Abbreviations: SP, Syntagmatic Paradigmatic; SET, String Edit Theory; RT, relational trace; EM, expectation maximization; LTM, long-term memory.

*E-mail: dennissj@psych.colorado.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0307758101

proposal of ref. 12, through theories that propose nine or 10 roles, such as ref. 13, to much more specific schemes that contain domain-specific slots, such as ORIG_CITY, DEST_CITY, or DEPART_TIME, that are used in practical dialogue understanding systems (14).

That there is much debate about semantic role sets and that existing systems must commit to a scheme *a priori* are important limitations of existing work and, I will argue, are consequences of a commitment to intentional semantics. In systems that use intentional semantics, the meanings of representations are defined by their intended use and have no inherent substructure.

For instance, the statement "Sampras outguns Agassi" might be represented as:

Sampras: Winner

Agassi: Loser

However, the names of the roles are completely arbitrary and carry representational content only by virtue of the inference system in which they are embedded.

Now contrast the above situation with an alternative extensional representation of "Sampras outguns Agassi," in which roles are defined by enumerating exemplars, as follows:

Sampras: Kuerten, Hewitt

Agassi: Roddick, Costa

The winner role is represented by the distributed pattern of Kuerten and Hewitt, words chosen because they are the names of people who have filled the "X" slot in a sentence like "X outguns Y" within the experience of the system. Similarly, Roddick and Costa are the names of people who have filled the "Y" slot in such a sentence and form a distributed representation of the loser role. Note the issue is not just a matter of distributed vs. symbolic representation. The tensor product representation used in the STAR model (15) of analogical reasoning uses distributed representations of the fillers but assigns a unique rank to each role and so is an intentional scheme. By contrast, the temporal binding mechanism proposed by ref. 16 allows for both distributed filler and role vectors and hence could implement extensional semantics.

The use of extensional semantics of this kind has a number of advantages. First, defining a mapping from raw sentences to extensional-meaning representations is much easier than defining a mapping to intentional representations, because it is now necessary only to align sentence exemplars from a corpus with the target sentence. The difficult task of either defining or inducing semantic categories is avoided.

Second, because the role is now represented by a distributed pattern, it is possible for the one-role vector to simultaneously represent roles at different levels of granularity. The pattern {Kuerten, Hewitt} could be thought of as a protoagent, an agent, a winner, and a winner of a tennis match, simultaneously. The role vectors can be determined from a corpus during processing, and no commitment to an *a priori* level of role description is necessary.

Third, extensional representations carry content by virtue of the other locations in the experience of the system where those symbols have occurred. That is, the systematic history of the comprehender grounds the representation. For instance, we might expect systematic overlap between the winner and person-who-is-wealthy roles, because some subset of {Kuerten, Hewitt} may also have occurred in an utterance such as "X is wealthy." These contingencies occur as a natural consequence of the causality being described by the corpus. We will call this type of implicit inference inference by coincidence and, as we will see in subsequent sections, the performance of the model is in large part due to this emergent property.

In the next section, we give a brief introduction to String Edit Theory (SET), which is used in the model to identify sentences from the corpus suitable for alignment with the current target and to define how these sentences should align. Next, the steps involved in interpreting a sentence in the model will be outlined. Then, the Tennis News domain that was chosen to test the model is described, and the results are presented. Finally, some factors that remain to be addressed are discussed.

Introduction to SET

SET was popularized in a book entitled *Time Warps, String Edits and Macromolecules* (17) and has been developed in both the fields of computer science and molecular biology (18–21). As the name suggests, the purpose of SET is to describe how one string, which could be composed of words, letters, amino acids, etc., can be edited to form a second string. That is, what components must be inserted, deleted, or changed to turn one string into another. In the model, SET is used to decide which sentences from a corpus are most like the target sentence, and which tokens within these sentences should align.

As an example, suppose we are trying to align the sentences "Sampras defeated Agassi" and "Kuerten defeated Roddick." The most obvious alignment is that which maps the two sentences to each other in a one-to-one fashion.

Sampras	defeated	Agassi	**[A1]**
Kuerten	defeated	Roddick	

In this alignment, we have three edit operations. There is a change of "Sampras" for "Kuerten," a match of "defeated," and a change of "Agassi" for "Roddick." In fact, this alignment can also be expressed as a sequence of edit operations,

⟨Sampras, Kuerten⟩
⟨defeated, defeated⟩
⟨Agassi, Roddick⟩

In SET, sentences do not have to be of the same length to be aligned. If we add "Pete" to the first sentence, we can use a delete to describe one way in which the resulting sentences could be aligned.

Pete	Sampras	defeated	Agassi	**[A2]**
–	Kuerten	defeated	Roddick	

The "–" symbol is used to fill the slot left by a deletion (or an insertion) and can be thought of as the empty word. The corresponding edit operation is denoted by ⟨Sampras, –⟩. Although these alignments may be the most obvious ones, there are many other options.

For instance, in aligning "Sampras defeated Agassi" and "Kuerten defeated Roddick," we could start by deleting "Sampras."

Sampras	defeated	Agassi	–	**[A3]**
–	Kuerten	defeated	Roddick	

Note that "Roddick" is now inserted at the end of the alignment (denoted ⟨–, Roddick⟩).

Alternatively, we could have deleted "Sampras" and then inserted "Kuerten," to give the following.

Sampras	–	defeated	Agassi	**[A4]**
–	Kuerten	defeated	Roddick	

There are a total of 63 ways in which "Sampras defeated Agassi" can be aligned with "Kuerten defeated Roddick," but not all of these alignments are equally likely. Intuitively, alignment **A4** seems better than **A3**, because the word "defeated" is matched. However, this alignment still seems worse than **A1** because it requires "Sampras" to be deleted and "Kuerten" to be inserted. A mechanism that produces alignments of sentences should favor those that have many matches and should penalize those that require many inser-

tions and deletions. To capture these intuitions, edit operations are assigned probabilities. Typically, match probabilities are higher than change probabilities, which are higher than insertion or deletion probabilities. Assuming conditional independence of the edit operations, the probability of an alignment is the multiplication of the probabilities of the edit operations of which it is comprised. Each alignment is an exclusive hypothesis about how the two strings might be aligned, and so the probability that the strings are aligned in one of these ways is the addition of the probabilities of the alignments. Given that there are an exponential number of alignments among strings, one may be concerned that any algorithm based on SET would be infeasible. However, there exist efficient dynamic programming algorithms that have O(nm) time and space complexity, where n and m are the lengths of the two strings (20).

Gap Probabilities

In the molecular biology literature, it is common to assign a lower probability to an initial insertion or deletion (known collectively as indels) and then higher probabilities to subsequent indels in the same block. As a consequence, alignments that involve long sequences of indels are favored over alignments that have many short sequences (22–24). In the context of macromolecule alignment, increasing the probability of block indels is desirable, because a single mutation event can often lead to a block insertion or deletion. An analogous argument is also applicable in the language case, because it is common for entire phrases or clauses to differentiate otherwise structurally similar sentences.

To illustrate the point, consider aligning the sentences "Sampras defeated Agassi" and "Sampras who defeated Roddick defeated Agassi." Two possible alignments are:

Sampras	–	–	–	defeated	Agassi
|	|	|	|	|	|
Sampras	who	defeated	Roddick	defeated	Agassi

[A5]

and

Sampras	–	defeated	–	–	Agassi
|	|	|	|	|	|
Sampras	who	defeated	Roddick	defeated	Agassi.

[A6]

Note that these alignments have the same matches and deletions and so will have the same probabilities as calculated above. However, Eq. **A5** should be preferred over Eq. **A6**, because it involves the block deletion of a clause. To capture this property, it is assumed that the probability of an initial indel is lower than the probability of a continuing indel. Now Eq. **A5** will be favored because it involves a single start indel and two subsequent indels, whereas Eq. **A6** has two start indels and one subsequent indel.† There exists an algorithm that calculates alignment probabilities under this model that retains O(nm) time and space complexity (22).

We have now completed the overview of SET. In the next section, the way in which the model exploits SET is described.

Description of the Syntagmatic Paradigmatic (SP) Model

In the SP model, sentence processing is characterized as the retrieval of associative constraints from sequential and relational long-term memory (LTM) and the resolution of these constraints in working memory. Sequential LTM contains the sentences from the corpus. Relational LTM contains the extensional representations of the same sentences (see Fig. 1).

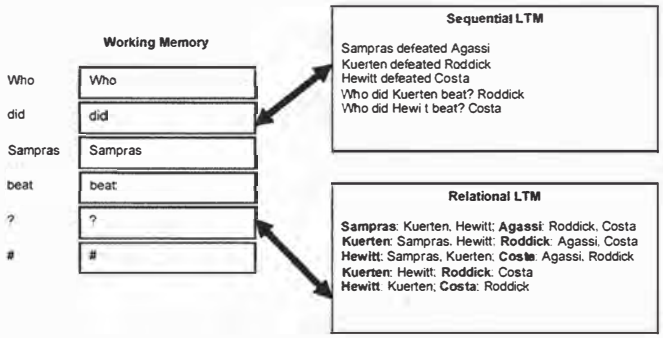


Fig. 1. SP architecture. #, empty slot. Ultimately, it will contain the answer to the question.

Creating an interpretation of a sentence/utterance involves the following steps.

Sequential Retrieval. The current sequence of input words is used to probe sequential memory for traces containing similar sequences of words. In the example, traces four and five, "Who did Kuerten beat? Roddick," and "Who did Hewitt beat? Costa," are the closest matches to the target sentence "Who did Sampras beat? #" and are assigned high probabilities (see Fig. 2).

To calculate the retrieval strength of a sequential trace, we take a similar approach to that adopted by the Bayesian models of recognition memory (25, 26), which have proven very successful at capturing a variety of memory effects.

Using the terminology $S_k \mapsto T$ to indicate that sequential trace S_k generated the target sentence T, we start with the odds ratio for $S_k \mapsto T$ given T and use the Bayes theorem to convert to a likelihood ratio:

$$\frac{P(S_k \mapsto T|T)}{P(\overline{S_k \mapsto T}|T)} = \frac{P(T|S_k \mapsto T)P(S_k \mapsto T)}{P(T|\overline{S_k \mapsto T})P(\overline{S_k \mapsto T})} = \frac{P(T|S_k \mapsto T)}{P(T|\overline{S_k \mapsto T})}, \quad [1]$$

assuming equal priors.

To calculate $P(T|S_k \mapsto T)$, we use as our data the possible edit sequences that may have been used to transform the trace into the target sentence. Each edit sequence represents an exclusive hypothesis about how S_k generated T, so the probabilities of these hypotheses can be added to determine $P(T|S_k \mapsto T)$:

$$P(T|S_k \mapsto T) = \sum_{a_p \in A_k} P(a_p|S_k \mapsto T), \quad [2]$$

where a_p is one of the edit sequences relating T and S_k, and A_k is the set of these sequences.

Assuming that the edit operations (match, change, insert, or delete) are sampled independently to create alignments:

$$P(T|S_k \mapsto T) = \sum_{a_p \in A_k} \prod_{e_r \in a_p} P(e_r|S_k \mapsto T), \quad [3]$$

where e_r is the rth edit operation in alignment a_p.

Similarly,

$$P(T|\overline{S_k \mapsto T}) = \sum_{a_p \in A_k} \prod_{e_r \in a_p} P(e_r|\overline{S_k \mapsto T}). \quad [4]$$

So, rearranging Eq. **1** and substituting in Eqs. **3** and **4**,

†Allison, Wallace, and Yee (21) describe this process in terms of a three-state finite-state machine and also generalize beyond the three-state case. Here, however, only the three-state case will be considered.

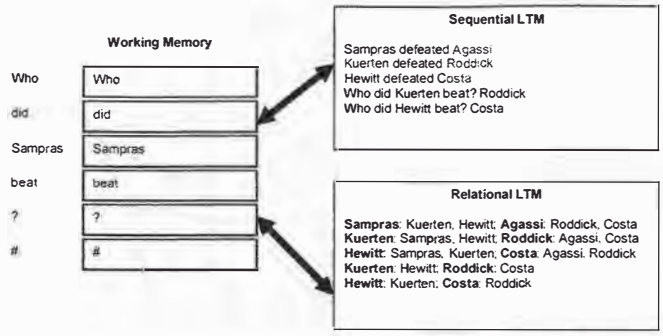


Fig. 2. Sequential retrieval. The traces "Who did Kuerten beat? Roddick" and "Who did Hewitt beat? Costa" are most similar to the input sentence "Who did Sampras beat? #" and are retrieved from sequential LTM. Bold type is used to indicate the traces that are retrieved.

$$P(S_k \mapsto T|T) = \frac{\sum_{a_p \in A_k} \prod_{e_r \in a_p} P(e_r|S_k \mapsto T)}{\sum_{a_p \in A_k} \prod_{e_r \in a_p} P(e_r|S_k \mapsto T) + \sum_{a_p \in A_k} \prod_{e_r \in a_p} P(e_r|S_k \mapsto T)}. \quad [5]$$

The expected retrieval probability is $P(S_k \mapsto T|T)$ normalized over the traces in memory,

$$\frac{P(S_k \mapsto T|T)}{\sum_j P(S_j \mapsto T|T)}.$$

Sequential Resolution. The retrieved sequences are then aligned with the target sentence to determine the appropriate set of substitutions for each word (see Fig. 3). Note that the slot adjacent to the "#" symbol contains the pattern {Costa, Roddick}. This pattern represents the role that the answer to the question must fill (i.e., the answer is the loser).

During sequential resolution, we are interested in calculating $E_k[P(\langle W_m, T_i\rangle|T)]$, the expected value of the probability that word T_i in slot i in the target sentence substitutes for the word W_m from the lexicon (W) in the context of sentence T.

We can write

$$E_k[P(\langle W_m, T_i\rangle|T)] = \sum_{k=1}^{N} \frac{P(S_k \mapsto T|T)}{\sum_{i=1}^{N} P(S_i \mapsto T|T)} P(\langle W_m, T_i\rangle|S_k \mapsto T, T), \quad [6]$$

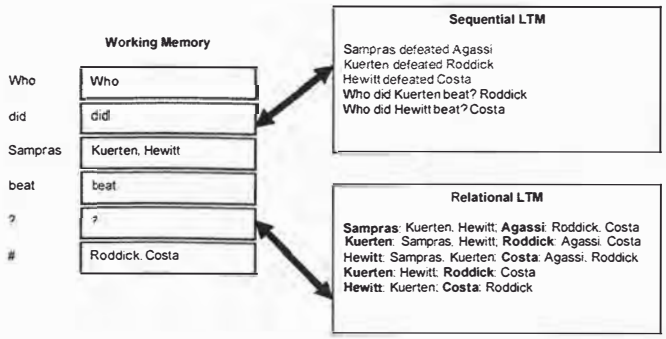


Fig. 3. Sequential resolution. Kuerten and Hewitt align with Sampras, and Roddick and Costa align with the answer slot ("#").

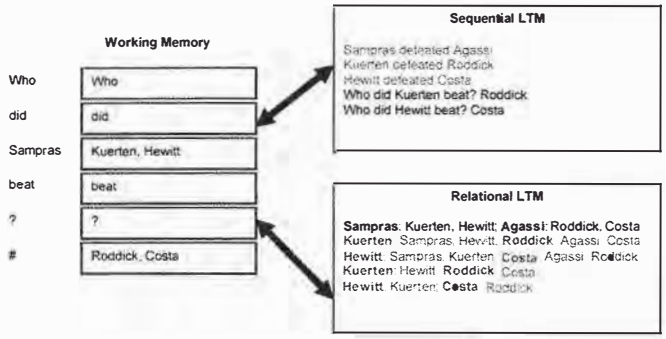


Fig. 4. Relational retrieval. The first relational trace is retrieved, because it contains similar role-filler bindings. Bold type is used to indicate the traces that are retrieved.

where N is the number of sequential traces in memory. Now we have divided the task into determining the probability that sequential trace k generated the target (which we calculated in the last section) and determining the probability that T_i and S_{kj} align given that trace k did generate the target.

Calculating the latter is straightforward, now that we have defined how alignments and edit operations are related. Because a given edit operation is either in an alignment or not, we can just add the probabilities of the alignments in which this change occurs and normalize by the probability of all the alignments:

$$P(\langle W_m, T_i\rangle|S_k \mapsto T, T) = \sum_{W_m = S_{kj}} P(\langle S_{kj}, T_i\rangle|S_k \mapsto T, T) = \sum_{W_m = S_{kj}} \frac{\sum_{\substack{a_p \in A_k \\ \langle S_{kj}, T_i\rangle \in a_p}} P(a_p|S_k \mapsto T)}{\sum_{a_p \in A_k} P(a_p|S_k \mapsto T)}. \quad [7]$$

We now have an algorithm with which we can calculate the probabilities of substitution within sentential context.

Relational Retrieval. The bindings of input words to their corresponding role vectors (the relational representation of the target sentence) are then used to probe relational LTM. In this case, trace one is favored because it involves similar role-filler bindings. That is, it contains a binding of Sampras onto the {Kuerten, Hewitt} pattern, and it also contains the {Roddick, Costa} pattern. Despite the fact that "Sampras defeated Agassi" has a different surface form than "Who did Sampras beat? #," it contains similar relational information and consequently has a high retrieval probability (Fig. 4).

As in the sequential case, when interpreting a new target sentence, we will assume that its relational trace (RT) has been generated via a set of edit operations on one of the RTs in memory. Specifically, we assume that each binding in RT, which we will denote RT_i, was generated by either an insert or by taking one of the bindings in the RT (R_{kj}) and editing the head word and role vector.

Applying the Bayes rule as we did in the sequential case, we get

$$\frac{P(R_k \mapsto RT|RT)}{P(R_k \mapsto RT|RT)} = \frac{P(RT|R_k \mapsto T)P(R_k \mapsto RT)}{P(RT|R_k \mapsto T)P(R_k \mapsto RT)}, = \frac{P(RT|R_k \mapsto RT)}{P(RT|R_k \mapsto RT)} \quad [8]$$

assuming equal priors. So we must now calculate $P(RT|R_k \mapsto RT)$. If RT contains M bindings, each of which are generated by independent operations,

$$P(RT|R_k \mapsto RT) = \prod_i^M P(RT_i|R_k \mapsto RT). \quad [9]$$

Furthermore, each binding in RT was generated from one of the bindings in R_k or by an insert, so

$$P(RT_i|R_k \mapsto RT) = \sum_j^N P(RT_i, R_{kj} \mapsto RT_i|R_k \mapsto RT) + P(RT_i, insert(RT_i)|R_k \mapsto RT), \quad [10]$$

and

$$\begin{aligned} P(RT|R_k \mapsto RT) &= \prod_i^M P(RT_i|R_k \mapsto RT) \\ &= \prod_i^M \left[\sum_j^N P(RT_i, R_{kj} \mapsto RT_i|R_k \mapsto RT) \right. \\ &\quad \left. + P(RT_i, insert(RT_i)|R_k \mapsto RT) \right] \\ &= \prod_i^M \left[\sum_j^N P(R_{kj} \mapsto RT_i|R_k \mapsto RT)P \right. \\ &\quad \left. \cdot(RT_i|R_{kj} \mapsto RT_i) + P(insert(RT_i)|R_k \mapsto RT) \right]. \end{aligned} \quad [11]$$

Note that if $R_{kj} \mapsto RT_i$, then $R_k \mapsto RT$, so $P(RT_i|R_{kj} \mapsto RT_i, R_k \mapsto RT) = P(RT_i|R_{kj} \mapsto RT_i)$. Also, $P(RT_i, insert(RT_i)|R_k \mapsto RT) = P(insert(RT_i)|R_k \mapsto RT)$. Now $P(RT_i|R_{kj} \mapsto RT_i)$ is the probability that the head word of RT_i (T_i) substitutes for the head word of R_{kj} (S_{kj}), and that the vector of change probabilities of RT_i is an edited version of the R_{kj} vector. To determine the probability of head-word substitution, the prior substitution probability can be used $P(\langle S_{kj}, T_i\rangle|S_k \mapsto T)$. To determine the probability of vector substitution, recall that each of the vectors is comprised of change probabilities. In each case, only one of the words could have substituted for their respective head words, so we can multiply the probability of the trace word (R_{kjl}) by the probability of the target word (RT_{ip}) and the probability that the trace word would substitute for the target word ($P(\langle W_p, W_l\rangle|S_k \mapsto T)$ to obtain an estimate of the probability that the role vector of R_{kj} was edited to produce the role vector of RT_i so

$$P(RT_i|R_{kj} \mapsto RT_i) = P(\langle S_{kj}, T_i\rangle|S_k \mapsto T)\sum_p \sum_l RT_{ip}P(\langle W_p, W_l\rangle|S_k \mapsto T)R_{kjl}, \quad [12]$$

where RT_{ip} is the pth component of the role vector of RT_i, and R_{kjl} is the lth component of the role vector of R_{kj}. The

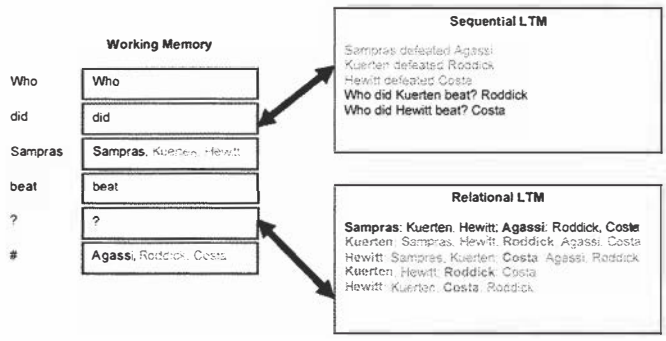


Fig. 5. Relational resolution. Agassi aligns with the answer slot, because it is bound to the {Roddick, Costa} pattern in the retrieved relational trace.

$P(RT_i|\overline{R_{kj} \mapsto RT_i})$, is calculated in an analogous way by using the $P(\langle S_{kj}, T_i\rangle|\overline{S_k \mapsto T})$ and $P(\langle W_p, W_l\rangle|\overline{S_k \mapsto T})$.

A similar logic is used to calculate the insertion probability

$$P(insert(RT_i)|R_k \mapsto RT) = P(\langle -, T_i\rangle|S_k \mapsto T)\sum_p P(\langle -, T_i\rangle|S_k \mapsto T)RT_{ip}. \quad [13]$$

And finally the retrieval component is

$$P(R_k \mapsto RT|RT) = \frac{P(RT|R_k \mapsto T)}{P(RT|R_k \mapsto T) + P(RT|\overline{R_k \mapsto T})}. \quad [14]$$

As before, the expected retrieval probability is $P(R_k \mapsto RT|RT)$ normalized over the traces in memory,

$$\frac{P(R_k \mapsto RT|RT)}{\sum_j P(R_j \mapsto RT|RT)}.$$

The above algorithm has constant space and time complexities $O(|T||S_k||W|^2)$, where $|W|$ is the size of the vocabulary. Although in principle this is expensive, in practice there is typically a small set of traces that attract the majority of the probability mass. Traces with very low retrieval probabilities are truncated and, as a consequence, there are usually only a few nonzero entries in each role vector.

Relational Resolution. Finally, the paradigmatic associations in the retrieved RTs are used to update working memory. In the RT for "Sampras defeated Agassi," "Agassi" is bound to the {Roddick, Costa} pattern. Consequently, there is a strong probability that "Agassi" should align with the "#" symbol, which, as a consequence of sequential retrieval, is also aligned with the {Roddick, Costa} pattern. Note that the model has now answered the question: it was Agassi who was beaten by Sampras (see Fig. 5).

As in the sequential case, we wish to calculate the probability that a given word substitutes for T_i given the relational representation of the target (RT).

$$E_k[P(\langle W_m, T_i\rangle|RT)] = \sum_{k=1}^N \frac{P(R_k \mapsto RT|RT)}{\sum_{i=1}^N P(R_i \mapsto RT|RT)} P(\langle W_m, T_i\rangle|R_k \mapsto RT, RT), \quad [15]$$

where $P(R_k \mapsto RT|RT)$ was calculated in the last section. To calculate the probability of substitution, we note that a substitution

of T_i for S_{kj} has occurred whenever $R_{kj} \mapsto RT_i$. As a consequence, the following derivation applies.

$$P(\langle W_m, T_i\rangle | R_k \mapsto RT, RT)$$
$$= \sum_{W_m = S_{kj}} P(\langle S_{kj}, T_i\rangle | R_k \mapsto RT, RT)$$
$$= \sum_{W_m = S_{kj}} \frac{P(RT_i | R_{kj} \mapsto RT_i)}{\sum_j P(RT_i | R_{kj} \mapsto RT_i) P(R_{kj} \mapsto RT_i | R_k \mapsto RT) + P(insert(RT_i | R_k \mapsto RT)} . \quad [16]$$

Combining Sequential and Relational Substitution Probabilities

We now have two procedures by which we can generate estimates of the substitution probabilities of trace and target words, one based on the sequence of words in the target (sequential) and one based on retrieved role-filler bindings (relational). The final question is how the estimates based on these two different sources of information should be combined to arrive at a final set of substitution probabilities. Taking a simple mixture of the information sources, we get:

$$P(\langle W_m, T_i\rangle) = \eta P(\langle W_m, T_i\rangle | T) + (1 - \eta) P(\langle W_m, T_i\rangle | RT), \quad [17]$$

where η is set at 0.5 for the simulations reported here.‡

To summarize, the model hypothesizes four basic steps. First, the series of words in the target sentence is used to retrieve traces that are similar from sequential LTM. Then, the retrieved sequential traces are aligned with the input sentence to create a relational interpretation of the sentence based on word order. This interpretation is then used to retrieve similar traces from relational LTM. Finally, working memory is updated to reflect the relational constraints retrieved in the previous step.

Updating Edit Probabilities with Corpus Statistics

The method used to derive the edit probabilities used above is a version of the Expectation Maximization (EM) algorithm (27). EM algorithms involve two steps. In the first step, the expected value of the log likelihood of the data given the current parameters is calculated. That is, we define Q:

$$Q(\theta, \theta^t) = \int_{y \in Y} \log(P(C, y | \theta) P(y | C, \theta^t) d\theta, \quad [18]$$

where C is the set of sentences in the training corpus, and Y is the set of all possible hidden variables (i.e., trace selections and edit sequences) that could have given rise to the set of traces. θ^t is the set of parameters at the current step.

In the second step, we find the parameters θ, which maximize Q. These will be used as the parameters for the next iteration of the algorithm

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta, \theta^t). \quad [19]$$

Repeated iterations of the EM algorithm are guaranteed to find a local minimum in the log likelihood (27). In the case of the SP model, the training algorithm reduces to adding the probabilities of the alignments in which each edit operation occurs and normalizing appropriately. Space precludes providing the entire derivation, but it follows the familiar pattern of EM derivations of mixture models (28).

Although the EM algorithm has proven useful in a wide range of language-learning tasks, optimization of the log likelihood of the data is not always a desirable objective (29). In the case of the SP model, a difficulty arises with the optimization of match probabilities. For low-frequency words, the probability that there will be a match of these words in the corpus can be very small, meaning that the match probabilities tend to zero. This property is particularly undesirable when the match probabilities are used in the relational model. For that reason, only change and indel probabilities were trained in the following evaluation.

The mathematical framework of the SP model has now been outlined. In the next section, we describe the data set used to test the question-answering capabilities of the model.

The Tennis News Domain

A number of criteria were used to select the domain on which to test the model. First, the domain was required to be one for which naturally occurring text was available, because it is important that the model be capable of dealing robustly with the variety of sentences typically found in real text. Also, in real corpora, there are many sentences that do not refer to the facts of interest at all, and the model should be capable of isolating the relevant ones.

Second, we wished to test the model's ability to extract relational information from sentences. Many question-answering systems use type heuristics rather than engaging in relational analysis. For instance, they might determine the date of the running of the Melbourne Cup by looking for sentences containing the term Melbourne Cup and returning any date within these sentences regardless of the role this date might fill. Although such heuristics are often very successful in practice, there are some questions for which a relational analysis is necessary.

Finally, we were interested in testing the model's ability to take advantage of inference by coincidence and so chose a domain in which the opportunities for such inferences are abundant.

Sixty-nine articles were taken from the Association of Tennis Professionals web site (www.atptennis.com). The articles were written between September 2002 and December 2002 and ranged in length from 134 to 701 words. In total, there are 21,212 words. The documents were manually divided into sentences, and the mean sentence length was 23.7.

The tennis domain fulfills each of the criteria. Naturally occurring text is available, and there were many nontarget sentences that the model was required to reject in its search for relevant information. Choosing the winner of a tennis match cannot be done by appealing to simple type heuristics, because relevant source sentences often contain the names of both the winner and the loser so that the correct answer must be selected from items of the same type. Finally, in sports reporting of this kind, there are often multiple cues, many of which are indirect, that allow the disambiguation of key facts, like who the winner of a match was.

Then 377 questions of the form "Who won the match between X and Y? X" were created. Any result that could be deduced from the article text was included. So, for instance, results that required the resolution of an anaphoric reference from other sentences in the same document were retained. Also, the winner was alternated between the first and second name positions so that the model could not simply repeat the name in the first slot to answer the question.

Results and Discussion

To test the model, the EM algorithm was first used to train edit probabilities and then each question was presented with the final answer slot vacant (e.g., "Who won the match between Sampras and Agassi? #"). The SP model was invoked to complete the pattern.§

‡I thank an anonymous reviewer who suggested using the explicit mixture model to combine sources.

§The parameters used were $P(\text{Match}) = 0.95$ $P(\text{Change}) = 0.025$ $P(\text{NotMatch}) = 0.35$, $P(\text{Not Change}) = 0.45$.

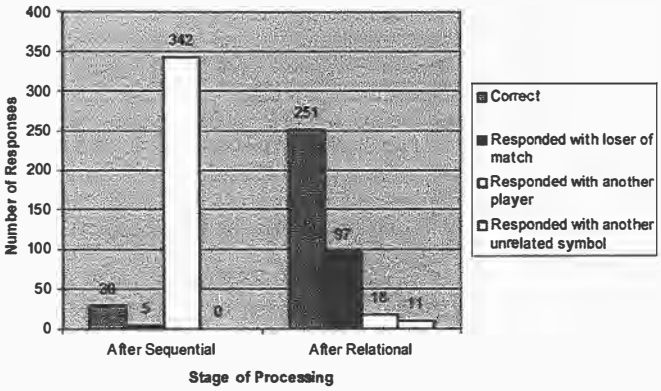


Fig. 6. Breakdown of result types after sequential and relational processing.

During sequential retrieval, the model was allowed to retrieve any sentence or question from the entire corpus. During relational retrieval, however, only facts derived from the sentences were allowed as retrieval candidates,[¶] that is, the factual knowledge embodied by the questions was not permitted to influence the results.

The token with the highest probability in the # slot was assumed to be the answer returned by the model. Fig. 6 shows a breakdown of the number of results in each category after sequential and relational resolution. After relational processing, on ≈67% of occasions the model correctly returned the winner of the match. Twenty-six percent of the time, it incorrectly produced the loser of the match. Five percent of the time, it responded with a player other than either the winner or loser of the match, and on 3% of occasions it committed a type error, responding with a word or punctuation symbol that was not a player's name.

There are a number of ways in which one might seek to establish an appropriate baseline against which to compare these results. Because the model is developed in a pattern-completion framework, it is possible for any symbol in the vocabulary to be returned. There were 2,522 distinct tokens in the corpus, so nominally the chance rate is <1%. However, one might also argue that the chance rate should be related to the number of elements of the appropriate type for a response, that is the number of names of players. There were 142 distinct players' names and so, by this analysis, the baseline would also be <1%. A further type distinction would be between winners and losers. There were 85 distinct winners, which results in a baseline of just >1%. Note that in any of these cases, the model is performing well above chance.

Another possible model for the decision process against which one might be tempted to compare the performance of the SP model is a winner maximum-likelihood model. In this model, the two players are extracted from the question, and the one that most often fills the winner slot is selected. With this model, performance is 74%. However, it is important to recognize that, to apply this model, one must provide a mechanism by which the relevant contenders are extracted from the sentence and which is capable of deciding what statistics are relevant for making frequency comparisons, decisions that will change on a question-by-question basis. By contrast, the SP model is only given a pattern to complete, and so is not only answering the question but is also extracting the relevant schema within which the question must be answered. In addition, when the SP model is run without relational retrieval or resolution, performance drops from 67% to 8% correct (see Fig. 6), so it would seem that relational processing was critical. Given that the questions were not included in relational memory, performance must have been driven by the statistics of the articles rather than of the questions. Consequently, the comparison against the maximum-likelihood model is somewhat inappropriate.

Issues That Compromised Performance

In examining the types of errors committed by the model, a number of recurring types were evidenced. As mentioned earlier, the use of anaphora is quite common in this corpus. The current model has no mechanism for the resolution of anaphora, which undermines its ability to both isolate the sentences containing the appropriate relational information and select the correct answer token. In addition, a mechanism for isolating appropriate context is necessary. On seven occasions in the current data set, there are sets of players for whom the questions are ambiguous without the use of context to isolate the correct match. In addition, inference by coincidence can sometimes induce an incorrect response. For instance, the model induces that Schalken won the match against Pete Sampras in part on the basis of the sentence "Schalken, from the Netherlands, made his best-ever grand slam showing at the US open last month. . ." However, although having a best-ever showing is indicative of winning, in this case, it is misleading because it was in fact Sampras who defeated Schalken in the semifinals. Finally, the model's lack of sensitivity to sublexical structure creates difficulties, particularly in deriving relational match when possessives are used. There are then many avenues by which one could look to improve performance.

Inference by Coincidence

To assess the contribution that inference by coincidence made to the performance of the model, the sentence with maximal retrieval probability for each query was classified into one of three categories.

The literal category contained those sentences where there was an explicit statement of the result, even if it required some interpretation. For example, when processing the question, "Who won the match between Ulihrach and Vicente? Ulihrach," the highest-probability relational trace was "Vicente bounced by Ulihrach," which literally states the result (even if it is necessary for one to interpret "bounced" in this context).

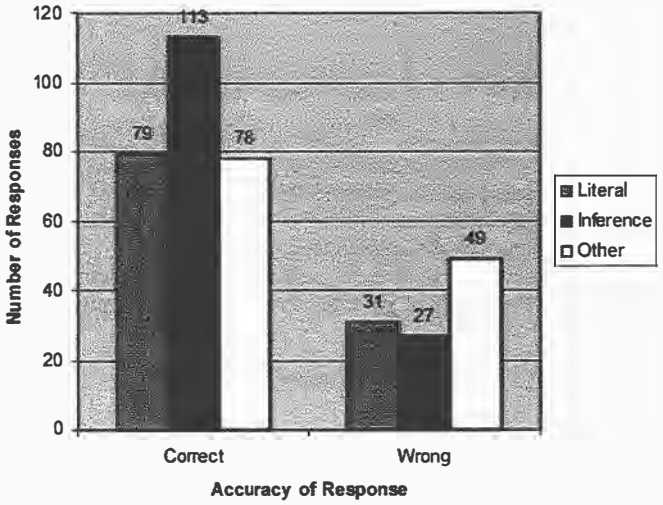


Fig. 7. Breakdown of responses based on the accuracy of the response and the type of the most probable relational trace according to the model. "Literal" refers to traces in which the answer was stated explicitly. "Inference" refers to traces in which the answer was not stated, but from which it could be inferred. "Other" refers to traces from which the answer was not derivable. Note that these statistics are, for the most part, probably trace only. The model, however, accumulates information from multiple traces, so it is still possible for it to answer correctly even if the most probable trace does not contain the relevant information.

[¶]To speed computation, only sentences that contained at least one of the two combatants were considered.

Table 1. Examples of inference by coincidence in the Tennis News domain

Who won the match between Carlsen and Kiefer? Carlsen.
Kafelnikov now meets Kenneth Carlsen of Denmark in the second round.
Who won the match between Kiefer and Safin? Safin.
Safin, Kafelnikov surge toward hometown showdown.
Who won the match between Ljubicic and Kutsenko? Ljubicic.
Sixth seed Davide Sanguinetti of Italy and eighth seed Ivan Ljubicic of Croatia took different paths to their opening-round wins at the president's cup in Tashkent.
Who won the match between Voltchkov and Haas? Voltchkov.
According to Haas, the injury first arose during Wednesday's match against Sargis Sargsian, and became progressively worse during practice and then the match against Voltchkov.
Who won the match between Srichaphan and Lapentti? Srichaphan.
Srichaphan has now won two titles in four finals this year.
Who won the match between Mamiit and Coria? Coria.
Kuerten, Coria withstand heat, set up fiery South American showdown.

Each example shows the question and the sentence that generated the most probable relational trace.

The inference category included those sentences that did not contain a literal statement of the result but that provided some evidence (not necessarily conclusive) for what the result may have been (see Table 1 for examples). For instance, when processing the question, "Who won the match between Portas and Sampras? Sampras," the relational trace with the highest retrieval probability was "Sampras claims 14th Grand Slam title." Although this sentence does not explicitly state the result of this match, one can infer that if Sampras won the title, then it is likely that he won this match. Note that this inference does not always follow, because the writer may have made reference to a result from a different tournament, or the question may have come from a different article. However, that Sampras won the title does provide evidence in favor of his having won this match. Unlike a traditional inference system, however, the SP model is making the inference by virtue of the fact that the names of people that appear in statements of the form "X claims title" also tend to appear in the winner slot at the end of the questions.

Finally, the other category included all remaining cases. These contained traces in which both players were mentioned but the sentence could not have been used to conclude who the winner may have been. For example, when the question, "Who won the match between Acasuso and Pavel? Acasuso" was presented, the most probable relational trace was "Pavel and Acasuso to clash in Bucharest semis." In addition, this category contains sentences that contradict the correct result. For example, the question "Who won the match between Pavel and Srichaphan? Pavel" produced the relational trace "Pavel, now 38-22 on the year, has reached two semifinals in 2002 Chennai I. to Srichaphan and Bucharest I. To Acasuso." This situation occurs when a player revenges an earlier loss. In addition, the other category was assigned when the sentence was unrelated to the question. For instance, when the model was presented with the question, "Who won the match between Meligeni and Etlis? Etlis," it returned "Kiefer quickly overcame Gaston Etlis of Argentina 6-2, 6-4 on Monday to qualify for the main draw of the Kremlin cup."

Fig. 7 shows the number of most probable relational traces in each category.

For an indication of the contribution that inference by coincidence is making to correct responding, consider those correct responses that can be attributed to either literal or inference traces. On 59% of occasions, the model was inferring the answer rather than relying on literal retrieval. Given that in each case a literal statement of the results existed in the corpus, it is significant that inference by coincidence seems to be playing such a crucial role in the performance of the model.

Conclusion

The ability of the SP model to isolate the combatants from arbitrary sentences and to successfully separate winners from losers demonstrates it is capable of extracting propositional information from text. Using simple retrieval and alignment operations, the model takes advantage of the statistics of word use. Unlike existing work (7, 8, 10), it need make no *a priori* commitment to particular grammars, heuristics, or sets of semantic roles, and it does not require an annotated corpus on which to train.

Furthermore, the large number of occasions (59%) on which the most probable relational trace was a sentence from which the result could be inferred but not directly derived is an indication that inference by coincidence can play a dominant role in successful question answering and may be a crucial factor in sentence comprehension in general.

I acknowledge the many discussions that have shaped the current work. In particular, I thank Michael Harrington, Michael Humphreys, Peter Kwantes, Andrew Smith, Walter Kintsch, Tom Landauer, Jose Quesada, and Michael Littman for helpful comments and suggestions. I also thank Jose Quesada for his assistance in creating the Tennis News data set. This research was supported by Australian Research Council Grant A00106012, U.S. National Science Foundation Grant EIA-0121201, and U.S. Department of Education Grant R305G020027.

1. Gaizauskas, R. & Wilks, Y. (1998) *J. Doc.* **54,** 70–105.
2. Cowie, J. & Lehnert, W. (1996) *Commun. ACM* **39,** 80–91.
3. Voorhees, E. M. (2002) *Eleventh Text Retrieval Conference (TREC 2002)*, eds. Voorhees, E. M. & Buckland, L. P., http://trec.nist.gov/pubs/trec11/t11_proceedings.html.
4. Moldovan, D., Harabagiu, S., Girju, R., Morarescu, P., Lacatusu, F., Novischi, A., Badulescu, A. & Bolohan, O. (2002) *Eleventh Text Retrieval Conference (TREC 2002)*, eds. Voorhees, E. M. & Buckland, L. P., http://trec.nist.gov/pubs/trec11/t11_proceedings.html.
5. Soubbotin, M. M. & Soubbotin, S. M. (2002) in *Eleventh Text Retrieval Conference (TREC 2002)*, eds. Voorhees, E. M. & Buckland, L. P., http://trec.nist.gov/pubs/trec11/t11_proceedings.html.
6. Fellbaum, C. (1998) WORDNET, *An Electronic Lexical Database* (MIT Press, Cambridge, MA).
7. Blaheta, D. & Charniak, E. (2000) in *Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics* (North Am. Chapter for the Assoc. for Computational Linguistics, Seattle), pp. 234–240.
8. Gildea, D. & Jurafsky, D. (2002) *Comput. Ling.* **28,** 245–288.
9. Palmer, M., Rosenzweig, J. & Cotton, S. (2001) *Proceedings of the First International Conference on Human Language Technology Research*, ed. Allan, J. (Morgan Kaufmann, San Francisco).
10. Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K. & Schasberger, B. (1993) *Comput. Ling.* **19,** 313–330.
11. Fillmore, C. J., Wooters, C. & Baker, C. F. (2001) in *Proceedings of the Pacific Asian Conference on Language, Information and Computation* (Pacific Asian Conference on Language, Information and Computation, Hong Kong).
12. Van Valin, R. D. (1993) *Advances in Role and Reference Grammar*, ed. Van Valin, R. D. (John Benjamins, Amsterdam).
13. Fillmore, C. J. (1971) in *22nd Round Table, Linguistics: Developments of the Sixties—Viewpoints of the Seventies*, ed. O'Brien, R. J. (Georgetown Univ. Press, Washington, DC), Vol. 24, pp. 35–56.
14. Stallard, D. (2000) in *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP'00)*, (citeseer.nj.nec.com/stallard00talkntravel.html), pp. 68–75.
15. Halford, G., Wilson, W., Guo, K., Gayler, R., Wiles, J. & Stewart, J. (1994) in *Analogical Connections*, eds. Holyoak, K. J. & Barnden, J. (Ablex, Norwood, MN), Vol. 2, pp. 363–415.
16. Hummel, J. & Holyoak, K. J. (1997) *Psychol. Rev.* **104,** 427–466.
17. Sankoff, D. & Kruskal, J. B. (1983) *Time Warps, String Edits and Macromolecules: The Theory and Practise of Sequence Comparison* (Addison–Wesley, New York).
18. Sellers, P. H. (1974) *J. Combin. Theor.* **16,** 253–258.
19. Levenshtein, V. I. (1965) *Dokl. Akad. Nauk. SSSR* **163,** 845–848.
20. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48,** 443–453.
21. Allison, L., Wallace, C. S. & Yee, C. N. (1992) *J. Mol. Evol.* **35,** 77–89.
22. Gotoh, O. (1982) *J. Mol. Biol.* **162,** 705–708.
23. Waterman, M. S., Smith, T. F. & Beyer, W. A. (1976) *Adv. Math.* **20,** 367–387.
24. Waterman, M. S. (1984) *Bull. Math. Biol.* **46,** 473–500.
25. Shiffrin, R. M. & Steyvers, M. (1997) *Psychon. Rev.* **4,** 145–166.
26. Dennis, S. & Humphreys, M. S. (2001) *Psychol. Rev.* **108,** 452–478.
27. Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) *J. R. Stat. Soc. B* **39,** 1–38.
28. Bilmes, J. A. (1998) Ph.D. thesis (University of California, Berkeley).
29. Klein, D. & Manning, C. D. (2001) in *Proceedings of the Conference on Computational Natural Language Learning*, eds. Daelemans, W. & Zajac, R. (Toulouse, France), pp. 113–120.

From paragraph to graph: Latent semantic analysis for information visualization

Thomas K. Landauer*†‡, Darrell Laham†, and Marcia Derr†

*Department of Psychology, University of Colorado, Boulder, CO 80309-0345; and †Knowledge Analysis Technologies, Boulder, CO 80301

Most techniques for relating textual information rely on intellectually created links such as author-chosen keywords and titles, authority indexing terms, or bibliographic citations. Similarity of the semantic content of whole documents, rather than just titles, abstracts, or overlap of keywords, offers an attractive alternative. Latent semantic analysis provides an effective dimension reduction method for the purpose that reflects synonymy and the sense of arbitrary word combinations. However, latent semantic analysis correlations with human text-to-text similarity judgments are often empirically highest at ≈300 dimensions. Thus, two- or three-dimensional visualizations are severely limited in what they can show, and the first and/or second automatically discovered principal component, or any three such for that matter, rarely capture all of the relations that might be of interest. It is our conjecture that linguistic meaning is intrinsically and irreducibly very high dimensional. Thus, some method to explore a high dimensional similarity space is needed. But the 2.7×10^7 projections and infinite rotations of, for example, a 300-dimensional pattern are impossible to examine. We suggest, however, that the use of a high dimensional dynamic viewer with an effective projection pursuit routine and user control, coupled with the exquisite abilities of the human visual system to extract information about objects and from moving patterns, can often succeed in discovering multiple revealing views that are missed by current computational algorithms. We show some examples of the use of latent semantic analysis to support such visualizations and offer views on future needs.

Most techniques for relating textual information rely on intellectually created links such as author-chosen keywords and titles, authority indexing terms, or bibliographic citations (1). Similarity of the semantic content of whole documents, rather than just titles, abstracts, or an overlap of keywords, offers an attractive alternative. Latent semantic analysis (LSA) provides an effective dimension reduction method for the purpose that reflects synonymy and the sense of arbitrary word combinations (2, 3).

Latent Semantic Analysis

LSA is one of a growing number of corpus-based techniques that employ statistical machine learning in text analysis. Other techniques include the generative models of Griffiths and Steyvers (4) and Erosheva *et al.* (5), and the string-edit-based method of S. Dennis (6) and several new computational realizations of LSA. Unfortunately, to date none of the other methods scales to text databases of the size often desired for visualization of domain knowledge. The linear singular value decomposition (SVD) technique described here has been applied to collections of as many as a half billion documents containing 750,000 unique word types, all of which are used in measuring the similarity of two documents. LSA presumes that the overall semantic content of a passage, such as a paragraph, abstract, or full coherent document, can be usefully approximated as a sum of the meaning of its words, as follows: meaning of paragraph ≈ meaning of word_1 + meaning of word_2 + . . . + meaning of word_n.

Mutually consistent meaning representations for words and passages can thus be derived from a large text corpus by treating each passage as a linear equation and the corpus as a system of simultaneous equations. In standard LSA, the solution of such a system is accomplished by SVD (3). SVD is defined as $X = WSP'$. As SVD is applied to a text corpus for LSA, X is a matrix of words by paragraphs, with cells containing the log of the frequency of a word in a paragraph weighted inversely with the entropy of the word across all paragraphs. The matrix is decomposed by an iterative sparse-matrix SVD program (3) into three matrices, two with orthonormal singular vectors, W and P, standing for words and paragraphs, respectively, and a diagonal S matrix of singular values (square roots of eigen values). SVD yields a solution that is unique up to linear transformation. For very large corpora, the methods find only approximate solutions in dimensionalities well below the rank of the matrix and, in practice, are usually limited to 200–400 dimensions, for reasons to be given shortly. Similarities between words or documents are usually measured by their cosines (cos) in the resulting high dimensional semantic space. Vectors for new paragraphs can be computed dynamically by simply adding the vectors of their words, although after large additions or changes in the domain, recomputing the semantic space may be necessary, a process that takes several hours to days depending on corpus size and computing power.

Even if mathematically provable, formal qualities such as resolution, compactness, and separation of clusters may not be what is most important for visualization (or other human uses such as information retrieval) unless they gave rise to useful human perceptions (or understandings). Therefore, we have tested the effectiveness of the underlying text analysis by simulating human judgments of the similarity of texts and comparing them with those of humans. This has been done in numerous ways with good results, agreement between machine and human being as good or almost as good as that between two humans. For example, after training on corpora from which humans learned or might have, LSA-based simulations have passed multiple choice vocabulary tests and textbook-based final exams at student levels (7). The frequently encountered effect of dimensionality and the existence of a high and strongly peaked optimum was dramatically shown by performance on multiple-choice items from the Test of English as a Foreign Language. LSA chose the most similar alternative word as that with the largest cos to the question word. Fig. 1 shows that its performance at 250–400 dimensions was very much better than at two or three,

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

Abbreviations: LSA, latent semantic analysis; SVD, singular value decomposition; MeSH, medical subject heading; cos, cosine.

‡To whom correspondence should be addressed. E-mail: landauer@psych.colorado.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0400341101

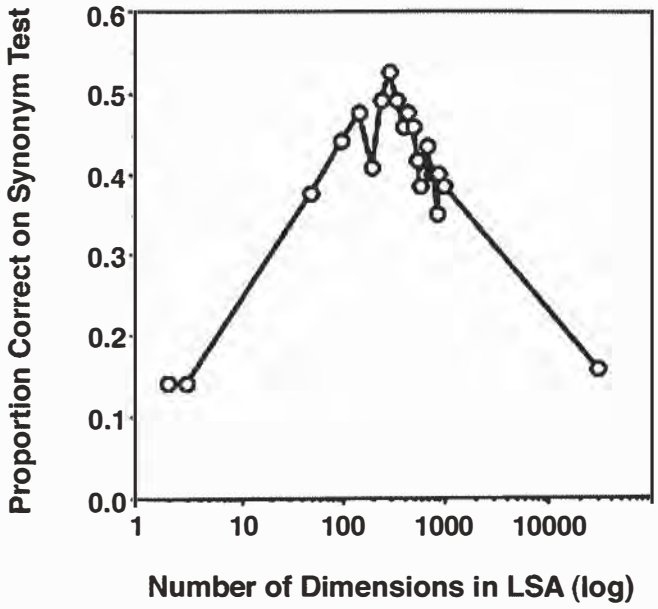


Fig. 1. LSA performance on the Test of English as a Foreign Language synonym test as a function of retained dimensions.

or at ones much higher than the optimum. At peak dimensionality, its score equaled that of successful applicants to U.S. colleges from non-English-speaking countries. In experiments simulating the amount and kind of reading of middle school students, LSA vocabulary growth equaled the average 10 per day increase for students (8). By matching documents with similar meanings but different words, LSA improved recall in information retrieval, usually achieving 10–30% better performance *cetera paribus* by standard metrics, again doing best with ≈300 dimensions (9). LSA has been found to measure coherence of text in such a way as to predict human comprehension as well as sophisticated psycholinguistic analysis, whereas measures of surface word overlap fail badly (10). By comparing contents, LSA predicted human ratings of the adequacy of content in expository test essays nearly as well as the scores of two human experts predicted each other, as measured by ≈90% as high mutual information between LSA and human scores as between two sets of human scores (7). The 300-dimension optimum is not a universal law, nor is there theory to explain it. The reason for finding it often (but not always) by empirical test is not known (note that result in scoring essays here relies exclusively on LSA as used in visualization and does not include other components used in automated essay scoring). To repeat, in our method, the measured relation between words is not the relative frequency with which they co-occur in the same documents, but the extent to which they have the same effect in the construction of total passage meanings. Nor is the relation between two paragraphs based on the literal words that they have in common, as in standard vector space information retrieval systems. Instead, it measures the extent to which the vectors of words they contain would add to form the same paragraph vectors independent of what sets of words went into the sum of either text. It is this property, and the empirical evidence that it produces representations that closely simulate human judgment, that is the basis of our belief that it offers important advantages for domain-knowledge representation. In particular, four relevant properties result for knowledge domain visualization purposes. (*i*) The method measures similarity of meaning of whole documents independent of the literal words used. For example, "the doctor operates on the patient" is highly similar to "the physician is in surgery" (cos = 0.86 ± 0.05; the standard deviation is based on random pairs from the same corpus) but considerably less similar to "a carpenter operates a saw patiently," which shares keywords but carries a completely different meaning (cos = 0.02 ± 0.05). (*ii*) It is sensitive to all similarities and differences between documents that are carried by word combinations, not just those of special interest or notice to their authors, other authors, or bibliographers. (*iii*) It ignores word order within documents and measures pairs of antonyms as equally similar to each other as pairs of synonyms (although the patterns of relations to other words of antonyms and synonyms, respectively, are quite different). These are disadvantages in some applications because nonlinear intrasentence syntactic and grammatical effects on meaning, such as predication, attachment, negation, and propositional implication, are lost ("no large proteins contain few amino acids" is very nearly the same, cos = 0.99, as "all amino acids contain many small proteins"). However, for most information retrieval and mapping purposes, ignoring these phenomena is of little consequence, or even advantageous. This is first because over paragraph and longer texts, their effects seem to be small, and, second, because measuring documents as closely related that assert different things about the same matters is usually desirable. (It would, of course, be useful if systems could also automatically detect significant differences in results and claims, in addition to topical similarity or "aboutness," but no general method by which this can be accomplished currently exists.) (*iv*) It is entirely automatic. It does not need human provision of key words or indexing, or even require that documents have been read, or aspects of their content noticed or appreciated by others. (Of course, citation offers additional information, for example, about influence, importance, and conceptual ancestry, but it is not always useful to confound these factors with content.)

Visualization Demonstrations

To be effective, an LSA representation of documents must start by deriving a good high dimensional "semantic space" for the whole domain or domains of knowledge to which the documents in question belong. As a rule of thumb, to attain adequate results, training data must include at least thousands of general or domain-relevant coherent passages, for which 75- to 125-word paragraphs are, empirically again, usually optimal. In our experience, the larger the training corpus the better, although there is some dependence on the size and homogeneity of the field to be covered and the size and specificity of its vocabulary. To achieve results as good as those described above, at least 200 dimensions usually must be retained.

For visualization, finding good projections is also important; many may be useful, most may not. There is no guarantee that any particular projection, the first and second principal components, or the 57th and 293rd, or the dynamically rotating 54th, 129th, and 200th, will reveal something familiar or new to the human visual system, or be of particular interest to a human analyst. It is also sometimes useful to compute a lower dimensional subspace representation of the relations between a small set of documents. For example, to map relations among a particular group of drugs, one might apply multidimensional scaling to a submatrix of cos to emphasize relations specific to the subset.

For visualizations, we have used the GGobi (11) high dimensional data viewer (see www.ggobi.org for current system reference and software). This system displays data points and lines in any subset of three dimensions passed to it as appropriately formatted files. It can apply sophisticated projection-pursuit hill climbing and randomization algorithms to automatically find dimension triplets and rotations to maximize properties such as dispersion or grouping of points. These may, of course, be useful for some purposes even without being comprehensible to a

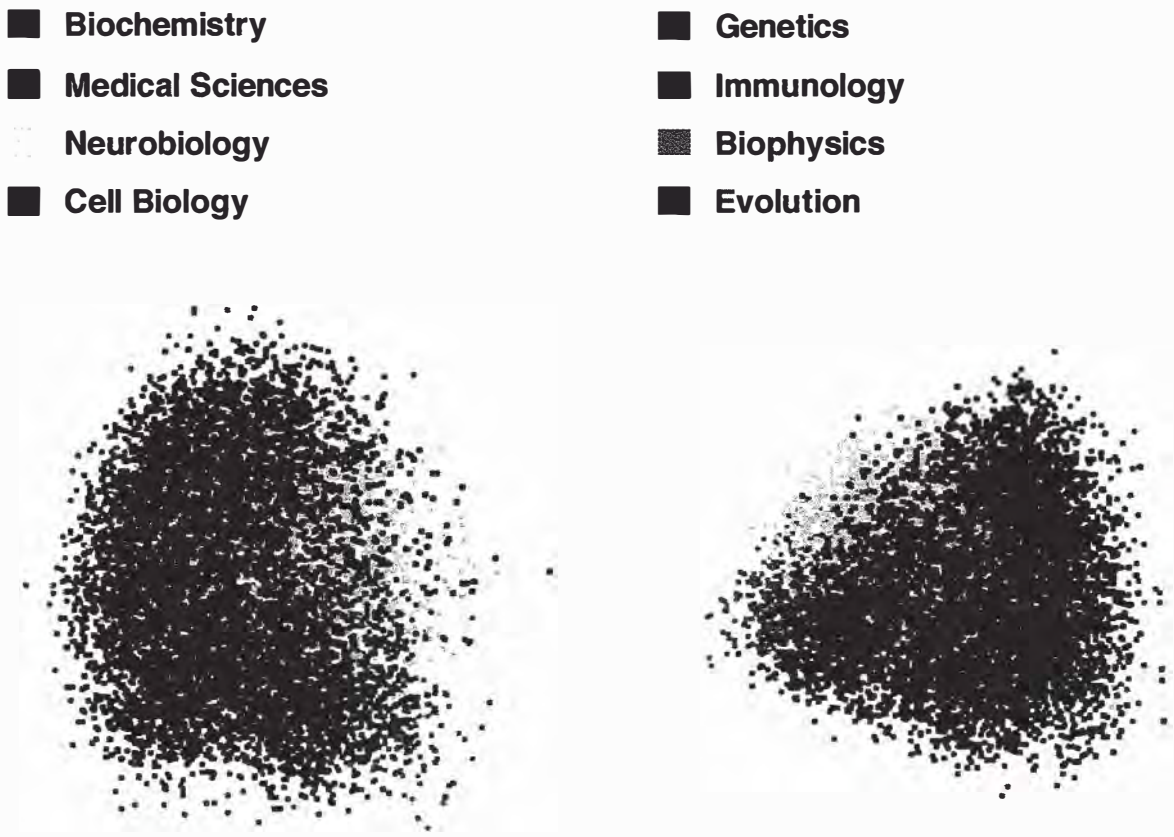


Fig. 2. PNAS articles colorized by biology subfield categories. The two-dimensional view on the three-dimensional space was selected algorithmically (*Left*) and by aided human selection (*Right*).

human observer (i.e., for "machine visualization"; e.g., of an irregular 10-dimensional solid). But our goal is human visualization, and we know of no way to assure value for that without bringing human observers into the picture. Automatic projection pursuit goes some way toward helping a human locate views of interest; however, (*i*) the process is heuristic and weak against the complexity of the space, and (*ii*) it does not necessarily correspond to the interests of human searchers. Thus, we believe that providing a kind of human–computer symbiosis in which the user can guide and evaluate what the system displays can add significant value. Importantly, GGobi allows a great deal of user control. Users can choose starting dimensions and control direction and speed of rotation. They can also specify the color and shape of glyphs for subsets of points and connect them with lines. Clicking on points can bring up associated text.

For the examples given here, we created a 300-dimensional LSA space from the full title, abstract, and body text of all articles in PNAS volumes 94–99, some 16,169 articles with a total of 67,341,938 word tokens containing 240,718 unique term types (no stemming or stop-listing was applied). In an expedience-dictated procedure differing from the optimal process described above, we first divided the corpus into 317,115 paragraph-like passages containing an average of 212 word tokens, and applied SVD to the resulting matrix of 240,718 terms by 317,115 passages. Having thus created a vector for each word type, we constructed vector representations of whole documents as the sum of their word vectors. (An alternative procedure would have been to construct article vectors from passage vectors, but that would have been inconsistent with the manner in which we added new documents not in the training corpus.)

We then used GGobi to search for revealing views. Data points for document sets of interest were visually identified by shape and/or coloring and are displayed in selected projections from among those examined. The views presented are ones we arrived at by the user-guided projection-pursuit methods described above, but for illustrative purposes restricted to dimensions 1–6. In the interest of consistency, comparison, and interpretability, in all of the views presented here we use the same generally good triple of dimensions, dimensions 3, 4, and 5 (which we found of greater interest than any other of the six-choose-three combinations, dimensions 1 and 2 in particular appearing to largely reflect word frequency, a dimension of little interest to us). It is worth noting that the "scree slope" in LSA decompositions is generally quite flat after the first two dimensions, each succeeding dimension contributing only a small and only slowly diminishing amount to the amount of total error reduction in the reconstructed X matrix. In each case, a combination of rapidly changing GGobi projection-pursuit views (which, of course, cannot be illustrated here) followed by more deliberate user control at interesting starting points was applied to find a better three-dimensional rotation, there being a virtually unlimited number. In every case, the selected rotation appeared more revealing to us than the initial one produced by the algorithm, the latter corresponding to the common use of two or three unrotated principal components.

Example visualizations of relations among the PNAS articles are shown in Figs. 2–6. Fig. 2 shows all articles from eight biology subfield categories in PNAS in the initial algorithmically chosen view (*Left*) and in our selected view (*Right*). This kind of display might aid in understanding the relations among nominal fields of science, or help editors, program managers, research organizations, or institutions organize publications, requests for proposals, or departments into maximally distinct and internally cohesive units. Used to display patterns over successive years, it could

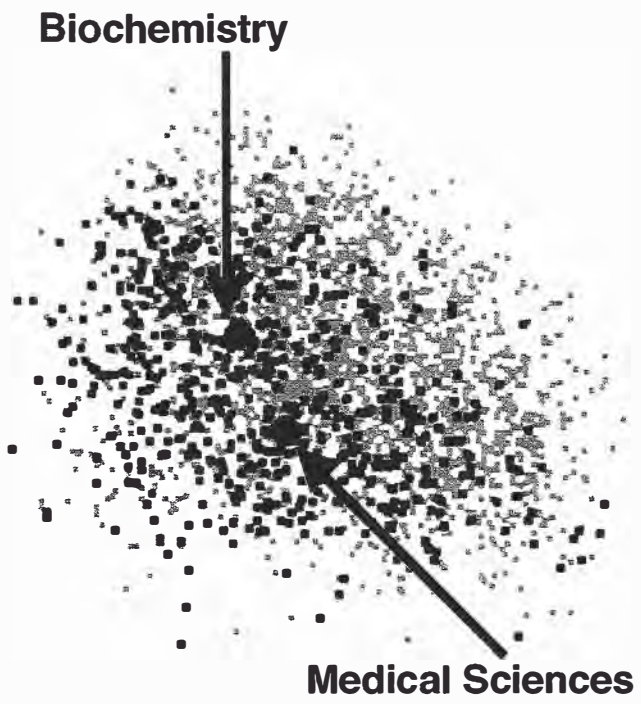


Fig. 3. Overlap of articles in categories Biochemistry (blue) and Medicine (red). Centroids of all articles in categories shown as the larger labeled dots.

aid analysis of the changing patterns of scientific effort. This case has special interest in that its value seems not to depend on identifying or characterizing individual documents.

Fig. 3 shows views of the overlap between two of those categories, Biochemistry (blue) and Medicine (red), with all of the rest displayed as gray dots. Also shown are centroids (the 300-dimensional average) of all documents in these two categories, respectively (the larger dots), and the category titles.

Fig. 4 illustrates a way to use the technique to find articles that relate two or more topics of interest in a particular way within the same article. They display documents similar to ones labeled with specific Medical Subject Headings (MeSH) terms, but not necessarily so labeled in the PNAS database, and do it in such a way that ones containing components of content of both kinds, either because they are the self-same article or because they coincide as seen from some particular point of view, stand out as new perceptual wholes. For this display, we first found all 721 PNAS94 articles with MeSH term "DNA" and all 1,401 with MeSH term "protein." For each set, we computed the centroid. We then found every article that had a cos $\geq$ 0.7 with each of the two centroids separately, that is, ones that contained a relatively high amount of relevant content. Those similar to the DNA centroid (but not necessarily so labeled in the PNAS database) are shown as open blue circles, and those similar to ones labeled "protein" are shown as orange dots. Ones $\geq$0.7 to both or coincident as seen from a particular viewpoint, appear as bull's-eyes of blue circles containing orange dots. Note that the blue and orange documents are not identified by MeSH terms but by their LSA similarities to the average content of articles with such terms. In fact, of the 397 with cos $\geq$ 0.7 to both DNA and protein centroids, fewer than half had both MeSH terms, and ones labeled with both did not appear as bull's-eyes in every view. The similarity threshold is continuously adjustable and need not be symmetrical. To us, moving the display through subsets of dimension triplets and rotating through three-dimensional viewing angles seems to have revealed patterns that are differentially interesting, whereas the first two principle components are less so. We hypothesize that for scientists expert in these overlapping fields, exploration of concentrated neighborhoods of bull's-eyes, clicking to see their titles or abstracts, or the less intrusive automatically generated keyword summaries, as shown here, could lead to useful information not as easily found by existing methods. Other variants are possible, for example precomputing bull's-eye documents and marking them so that they appear in every view. No discrete divisions or boundary planes, which are almost always artificial, are computed, and the natural fuzziness and intermingling is carried into the display. What is shown is a complete picture of how objects from different fuzzy classes are distributed with respect to themselves and each other from one perspective selected for its utility to the user with the help of the computational algorithm.

To compare this visualization to a completely verbal presentation of the same data, we computed joint topicality by multiplying the cos to the two centroids for each bull's-eye article. Table 1 shows the top two and the bottom two article titles in amount of overlap among the 397 articles, along with the product of the two cos.

Intuitively, it seems that the verbal presentation offers more precise information for choosing cases to examine, whereas the visual presentation offers a more flexible style of exploration that better shows multiple, fuzzy, and intermixed and complexly patterned relations among the documents. In addition, note that to explore the relations from a different perspective (similar to a different facet in information retrieval terminology), a whole new relevance ranked list would have to be produced and examined.

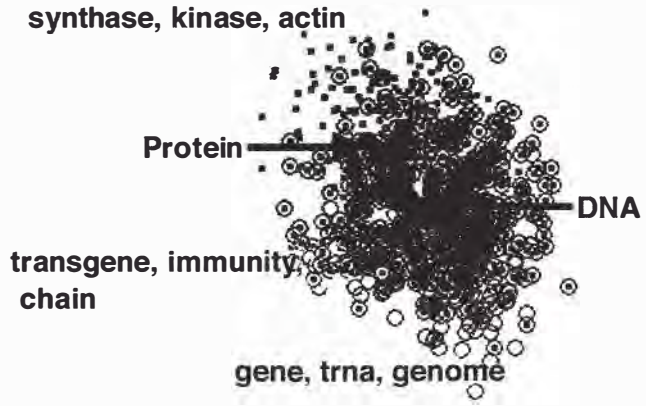


Fig. 4. Overlap between articles similar at cos $\geq$ 0.7 to centroids of ones with MeSH terms DNA or protein. Note the groups of bull's-eyes, articles related to both topics according in the current view, and autogenerated key words.

Table 1. Top two and bottom two titles in the amount of topic overlap, as determined by cos product

Title	cos product
In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition	0.664
Prospero is a panneural transcription factor that modulates homeodomain protein activity	0.656
Chondrocytes as a specific target of ectopic Fos expression in early development	0.492
c-Myc transactivation of LDH-A: Implications for tumor metabolism and growth	0.490

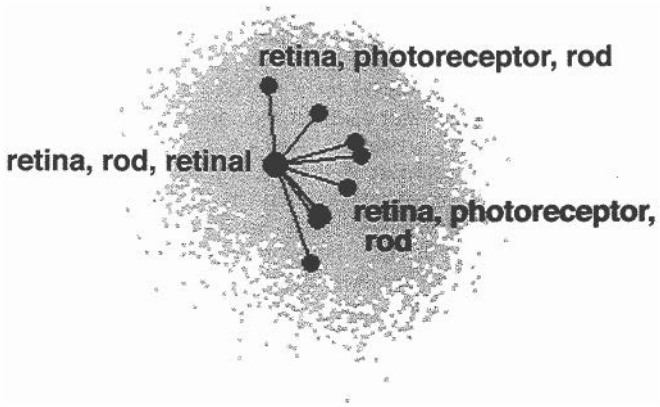


Fig. 5. Connecting similar articles across years. Red, a single article from 1998; green, similar ones from 1997; and blue, similar ones from 1999, labeled by autogenerated key words.

For Fig. 5, we started with one 1998 article and connected it to the most similar articles from 1997 and 1999. Investigations of this kind might help a researcher search for precedents, background work, or new related work before the sometimes considerable delay before indexing and citations provide good coverage. A historian might look for lines of progress, independent discoveries, or missed opportunities. The apparent advantage over corresponding lists would be the visualization of the mutual distributions (in multiple dimensions). A disadvantage, again, is in the more awkward identification of the actual articles.

Fig. 6 shows an application more similar to traditional query-based information retrieval. In Fig. 6*A*, the dispersion of all documents (full text of whole document) in 6 years of PNAS articles is shown in SVD dimensions 3 and 4. In all of these figures, the article title *Primordial nucleosynthesis* was used as the query (shown as the black circle).

Red, green, and blue circles, respectively, indicate the position of documents whose similarity to the query is more than four, three, and two standard deviations above the mean similarity of randomly chosen documents. Because human similarity judgments are monotonic with LSA cos, one can say that less than one in a thousand documents would be judged to be at least as similar in meaning to the query as the red documents.

An additional interesting feature of such views is the extent of mixture of relevant and nonrelevant articles and the differential patterning of closeness to the query in different directions. Such patterns also, of course, would vary with the choice of viewing plane. This illustrates well the loss of potentially interesting detail in standard ranked return lists (and purely algorithmic choice of view). In Figs. 6 *B–D*, closeness to the query is assigned to the dimension orthogonal to the plane of Fig. 6*A*, and the

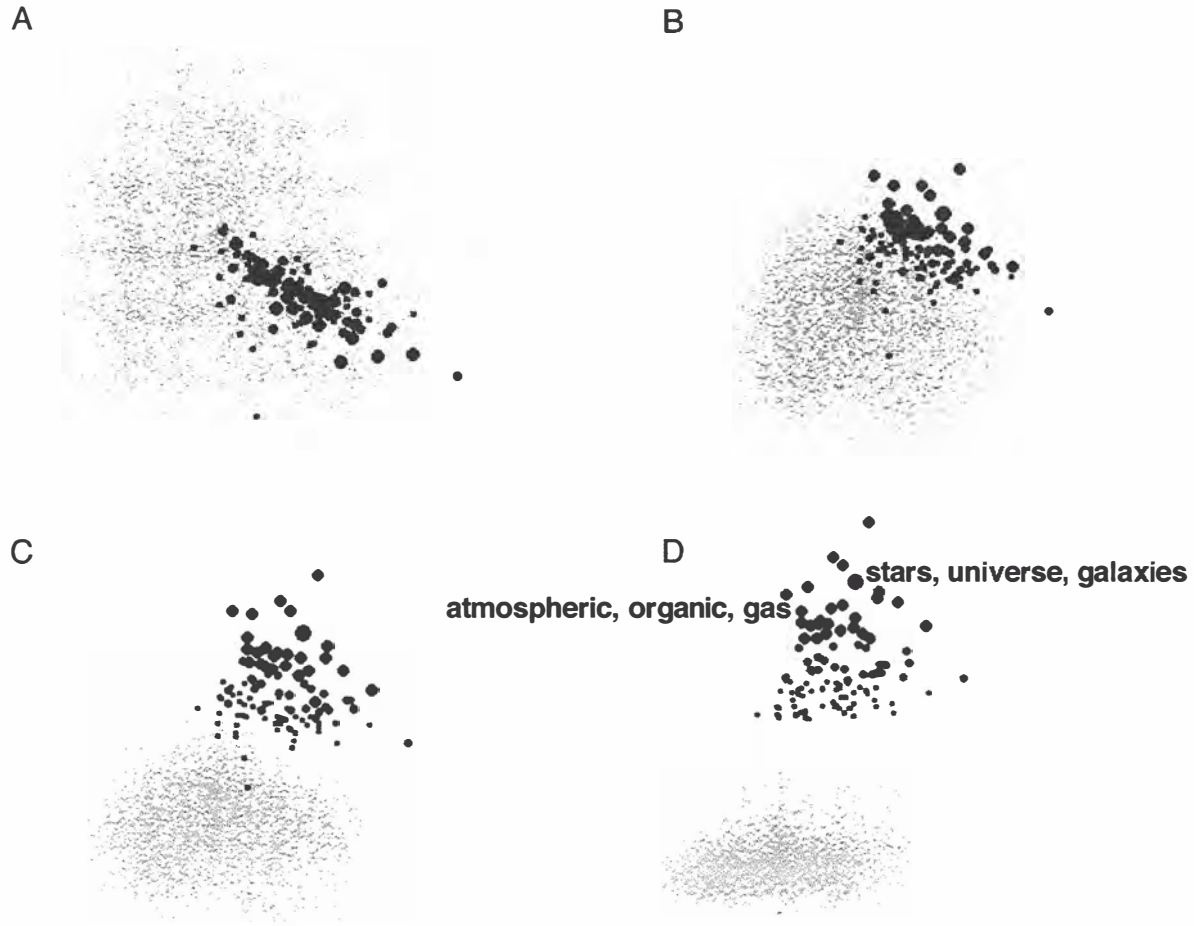


Fig. 6. Rotation from knowledge map view *A* (SVD dimensions 3 and 4), through views *B* and *C*, to information retrieval view *D* (SVD dimension 3 and relevance to query). Query is marked as a black dot; significant results are marked in red, green, and blue (see text).

plane rotated by hand and eye to two particularly interesting views, one that stretches the orthogonal relevance view to emphasize the distribution of relevances, the other to spread the relevance peak to reveal qualitatively different ways in which similar relevances are attained. In the last frame, two articles from apparently different neighborhoods are automatically labeled to show that they indeed appear to be topically distinct; presumably, the user would choose to examine only articles in one of the neighborhoods. This kind of search has much the same goal as recent attempts to automatically cluster subsets of returns, but allows and relies on visual search by the user that can reveal patterns and shapes such as clouds, gradual intermixings, and scattered islands that hard-boundary clustering usually misses, but the human visual system has evolved to perform in still mysterious ways.

Of course, it might sometime be possible to find computational procedures to automatically find views optimized according to these or other objectives. If and when user control is better is an open question. Another open question is the extent to which such visual explorations are useful, given the greatly impoverished semantic information they carry; without the labels, the reader has no idea what the visualizations are about. Labels for more than a few of the dozens or hundreds of points obscure the rest. How much help are the clouds of unlabeled points and for what tasks? How much help are short labels? Would interpretability of dimensions help? LSA dimensions as extracted are fundamentally uninterpretable because of the indeterminacy of rotation. However, it should be possible to label them dynamically in the same rather minimal manner as we have labeled points.

Two More Examples of Potential Applications

There are a very large number of possible ways to use semantic content-based measures of similarity in visualization, most of which we have probably not yet imagined, and space prevents showing more of the ones we have. However, to suggest the possible range, here are just two more ideas. (*i*) One could display the multidimensional topical distribution of research proposals in one color and that of bibliographies from vitae of potential reviewers in another, and explore views to find a minimal set of reviewers whose expertise best covers the subjects of all of the proposals. (*ii*) One could plot each article from a large number of journals in two dimensions, differentially coloring them and labeling centroid points with, say, the number of citations to each weighted by the cos of citing to cited, as a third dimension. Narrow, coherent fields would stand out as high pointed peaks, related groups as ranges separated by valleys. This would be a semantic full-content version of something that has frequently been done with other kinds of linkages.

Comments on the Enterprise

First, we conjecture that verbal meaning is irreducibly high dimensional. Thus, the value of automatic reductions to two or three best dimensions may be inherently limited; although they may be valuable for some purposes, they must often provide only an impoverished and possibly misleading impression of the relations in a dataset. Different researchers and scholars are often interested in different aspects of articles, only some of which may have been indexed, key-worded, the object of citation, or shown in a particular view. The alternative we have explored here is a combination of measuring similarity of the entire content of articles with high dimensional visualizations that support search for projections that are of special interest to the user. Our goal in selecting examples has been to identify cases that exploit the putative advantages of these approaches. Unfortunately, we do not know whether we have succeeded because we have not yet tested typical users using the displays to perform either typical or novel tasks. This appears to be a widespread situation in research in information visualization. Seldom have new visual displays been empirically compared with best-of-class verbal methods for the same tasks. The consequence is that the majority of work in the field is, like ours, technology driven rather than user problem driven and user success tested.

Despite decades of highly creative and sophisticated innovation, and a plethora of claims for obvious superiority of the visualization approach, we do not see visual maps of verbal information in popular and effective use. It is, of course, possible that visualizing verbal information is in large part just an appealing bad idea. A more optimistic view is that the application of more user testing to understand what does and doesn't help people do what, will steer innovations in more effective directions. Precedent exists, for example in Bellcore's *SuperBook*, for turning novel information search devices from useless as first designed to order-of-magnitude more effective through iterative empirical usability analysis and redesign (12).

This work was supported by the National Science Foundation, Air Force Research Laboratories, U.S. Army Research Institute for the Behavioral and Social Sciences, and the Office of Naval Research.

1. Börner, K., Chen, C. & Boyack, K. W. (2003) *Annu. Rev. Info. Sci. Technol.* **37,** 179–255.
2. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990) *J. Am. Soc. Info. Sci.* **41,** 391–407.
3. Berry, M. W. (1992) *Int. J. Supercomputer Appl.* **6,** 13–49.
4. Griffiths, T. L. & Steyvers, M. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5228–5235.
5. Erosheva, E., Fienberg, S. & Lafferty, J. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5220–5227.
6. Dennis, S. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5206–5213.
7. Landauer, T. K., Foltz, P. & Laham, D. (1998) *Discourse Processes* **25,** 259–284.
8. Landauer, T. K. & Dumais, S. T. (1997) *Psychol. Rev.* **104,** 211–240.
9. Dumais, S. T. (1994) in *The Second Text Retrieval Conference (TREC2)*, ed. Harman, D. (Natl. Inst. Stand. Technol., Gaithersburg, MD), pp. 105-116.
10. Landauer, T. K., Laham, D. & Foltz, P. W. (2003) in *Automated Essay Scoring: A Cross-Disciplinary Perspective*, eds. Shermis, M. D. & Burstein, J. (Lawrence Erlbaum Associates, Mahwah, NJ), pp. 87–112.
11. Swayne, D. F., Cook, D. & Buja, A. (1998) *J. Comput. Graphical Stat.* **7,** 113-130.
12. Egan, D. E., Remde, J. R., Gomez, L. M., Landauer, T. K., Eberhart, J. & Lochbaum, C. C. (1989) *ACM Trans. Info. Systems* **7,** 30–57.

Mixed-membership models of scientific publications

Elena Erosheva*†, Stephen Fienberg‡§, and John Lafferty§¶

*Department of Statistics, School of Social Work, and Center for Statistics and the Social Sciences, University of Washington, Seattle, WA 98195; and ‡Department of Statistics, ¶Computer Science Department, and §Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA 15213

PNAS is one of world's most cited multidisciplinary scientific journals. The PNAS official classification structure of subjects is reflected in topic labels submitted by the authors of articles, largely related to traditionally established disciplines. These include broad field classifications into physical sciences, biological sciences, social sciences, and further subtopic classifications within the fields. Focusing on biological sciences, we explore an internal soft-classification structure of articles based only on semantic decompositions of abstracts and bibliographies and compare it with the formal discipline classifications. Our model assumes that there is a fixed number of internal categories, each characterized by multinomial distributions over words (in abstracts) and references (in bibliographies). Soft classification for each article is based on proportions of the article's content coming from each category. We discuss the appropriateness of the model for the PNAS database as well as other features of the data relevant to soft classification.

> The Proceedings is there to help bring new ideas promptly into play. New ideas may not always be right, but their prominent presence can lead to correction. We must be careful not to censor even those ideas which seem to be off beat.
>
> Saunders MacLane (1)

Are there internal categories of articles in PNAS that we can obtain empirically with statistical data-mining tools based only on semantic decompositions of words and references used? Can we identify MacLane's "off-beat" but potentially path-breaking PNAS articles by using these internal categories? Do these empirically defined categories correspond in some natural way to the classification by field used to organize the articles for publication, or does PNAS publish substantial numbers of interdisciplinary articles that transcend these disciplinary boundaries? These are examples of questions that our contribution to the mapping of knowledge domains represented by PNAS explores.

Mathematical and statistical techniques have been developed for analyzing complex data in ways that could reveal underlying data patterns through some form of classification. Computational advances have made some of these techniques extremely popular in recent years. For example, 2 of the 10 most cited articles from 1997–2001 PNAS publications are on applications of clustering for gene-expression patterns (2, 3). The traditional assumption in most methods that aim to discover knowledge in underlying data patterns has been that each subject (object or individual) from the population of interest inherently belongs to only one of the underlying subpopulations (clusters, classes, aspects, or pure type categories). This implies that a subject shares all its attributes, usually with some degree of uncertainty, with the subpopulation to which it belongs. Given that a relatively small number of subpopulations is often necessary for a meaningful interpretation of the underlying patterns, many data collections do not conform with the traditional assumption. Subjects in such populations may combine attributes from several subpopulations simultaneously. In other words, they may have a mixed collection of attributes originating from more than one subpopulation.

Several different disciplines have developed approaches that have a common statistical structure that we refer to as mixed membership. In genetics, mixed-membership models can account for the fact that individual genotypes may come from different subpopulations according to (unknown) proportions of an individual's ancestry. Rosenberg *et al.* (4) use such a model to analyze genetic samples from 52 human populations around the globe, identifying major genetic clusters without using the geographic information about the origins of individuals. In the social sciences, such models are natural, because members of a society can exhibit mixed membership with respect to the underlying social or health groups for a particular problem being studied. Hence, individual responses to a series of questions may have mixed origins. Woodbury *et al.* (5) use this idea to develop medical classification. In text analysis and information retrieval, mixed-membership models have been used to account for different topical aspects of individual documents.

In the next section, we describe a class of mixed-membership models that unifies existing special cases (6). We then explain how this class of models can be adapted to analyze both the semantic content of a document and its citations of other publications. We fit this document-oriented mixed-membership model to a subcollection of the PNAS database supplied to the participants in the Arthur M. Sackler Colloquium Mapping Knowledge Domains. We focus in our analysis on a high-level description of the fields in biological sciences in terms of a small number of extreme or basis categories. Griffiths and Steyvers (7) use a related version of the model for abstracts only and attempt a finer level of description.

Mixed-Membership Models

The general mixed-membership model that we work with relies on four levels of assumptions: population, subject, latent variable, and sampling scheme. Population level assumptions describe the general structure of the population that is common to all subjects. Subject-level assumptions specify the distribution of observable responses given individual membership scores. Membership scores are usually unknown and hence can be viewed also as latent variables. The next assumption is whether the membership scores are treated as fixed or random in the model. Finally, the last level of assumptions specifies the number of distinct observed characteristics (attributes) and the number of replications for each characteristic. We describe each set of assumptions formally in turn.

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

†To whom correspondence should be addressed. E-mail: elena@stat.washington.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0307760101

Population Level. Assume there are K original or basis subpopulations in the populations of interest. For each subpopulation k, denote by $f(x_j|\theta_{kj})$ the probability distribution for response variable j, where θ_{kj} is a vector of parameters. Assume that, within a subpopulation, responses to observed variables are independent.

Subject Level. For each subject, membership vector $\lambda = (\lambda_1, \ldots, \lambda_K)$ provides the degrees of a subject's membership in each of the subpopulations. The probability distribution of observed responses x_j for each subject is defined fully by the conditional probability $Pr(x_j|\lambda) = \sum_k \lambda_k f(x_j|\theta_{kj})$ and the assumption that response variables x_j are independent, conditional on membership scores. In addition, given the membership scores, observed responses from different subjects are independent.

Latent-Variable Level. With respect to the latent variables, one could assume that they are either fixed unknown constants or random realizations from some underlying distribution.

1. If the membership scores λ are fixed but unknown, the conditional probability of observing x_j, given the parameters $\boldsymbol{\theta}$ and membership scores, is

$$Pr(x_j|\lambda;\ \boldsymbol{\theta}) = \sum_{k=1}^{K} \lambda_k f(x_j|\theta_{kj}). \qquad [1]$$

2. If membership scores λ are realizations of latent variables from some distribution D_α, parameterized by vector α, then the probability of observing x_j, given the parameters, is

$$Pr(x_j|\alpha,\ \boldsymbol{\theta}) = \int \left(\sum_{k=1}^{K} \lambda_k f(x_j|\theta_{kj}) \right) dD_\alpha(\lambda). \qquad [2]$$

Sampling Scheme. Suppose R independent replications of J distinct characteristics are observed for one subject, $\{x_1^{(r)}, \ldots, x_J^{(r)}\}_R^{r=1}$. Then, if the membership scores are treated as realizations from distribution D_α, the conditional probability is

$$Pr\left(\{x_1^{(r)}, \ldots, x_J^{(r)}\}_{r=1}^{R} \,\middle|\, \alpha,\ \boldsymbol{\theta} \right) = \int \left(\prod_{j=1}^{J} \prod_{r=1}^{R} \sum_{k=1}^{K} \lambda_k f(x_j^{(r)}|\theta_{kj}) \right) dD_\alpha(\lambda). \qquad [3]$$

When the latent variables are treated as unknown constants, the conditional probability for observing R replications of J variables can be derived analogously. In general, the number of observed characteristics J does not need to be the same across subjects, and the number of replications R does not need to be the same across observed characteristics.

One can derive examples of mixed-membership models from this general set up by specifying different choices of J and R and different latent-variable assumptions. Thus, the "grade-of-membership" model of Manton *et al.* (8) assumes that polytomous responses are observed to J survey questions without replications and uses the fixed-effects assumption for the membership scores. Potthoff *et al.* (9) use a variation of the grade-of-membership model by treating the membership scores as Dirichlet random variables; the authors refer to the resulting model as "Dirichlet generalization of latent class models." Erosheva (6) provides a formal latent-class representation for the grade-of-membership model approach. In genetics, Pritchard *et al.* (10) use a clustering model with admixture. For diploid individuals, the clustering model assumes that $R = 2$ replications (genotypes) are observed at J distinct locations (loci), treating the proportions of a subject's genome that originated from each of the basis subpopulations as random Dirichlet realizations. Variations of mixed-membership models for text documents called "probabilistic latent semantic analysis" (11) and "latent Dirichlet allocation" (12) both assume that a single characteristic (word) is observed a number of times for each document, but the former model considers the membership scores as fixed unknown constants, whereas the latter treats them as random Dirichlet realizations.

The mixed-membership model framework presented above unifies several specialized models that have been developed independently in the social sciences, genetics, and text-mining applications. In the text-mining area, initial work by Hofmann (11) on probabilistic latent semantic analysis was followed by the work of Blei *et al.* (12), who proposed a Dirichlet generating distribution for the membership scores and the use of variational methods to estimate the latent Dirichlet allocation model parameters. Minka and Lafferty (13) developed a more accurate approximation method for this model.

A natural extension of the original analyses in the text-mining area that have been based on a single source is to combine information from multiple sources. Cohn and Hofmann (14) propose a probabilistic model of document content and hypertext connectivity for text documents by considering links (or references) in addition to words, thus essentially combining two distinct characteristics; they treat the membership scores as fixed. Following Cohn and Hofmann, we adopt a mixed-membership model for words and references in journal publications but treat the membership scores as random Dirichlet realizations. Barnard *et al.* (15) develop similar and alternative approaches for combining different sources of information.

Mixed-Membership Models for Documents

We can use the general model framework for documents consisting of abstracts and references by representing a document as $d = (\{x_1^{(r_1)}\}, \{x_2^{(r_2)}\})$, where $x_1^{(r_1)}$ is a word (w) in the abstract and $x_2^{(r_2)}$ is a reference (r) in the bibliography, $r_j = 1, \ldots, R_j$. By adopting the "bag-of-words" assumption, we treat the words in each abstract as independent replications of the first observed characteristic (word). Similarly, under the assumption of a "bag of references," we treat references as independent replications of the second observed characteristic (reference). Thus, the representation of a document consists of word counts $n(w, d)$ (the number of times word w appears in document d) and reference counts $n(r, d)$ (1 if the bibliography of d contains a reference to r, and 0 otherwise). In this context, subpopulations refer to topical aspects.

The parameters $\boldsymbol{\theta}$ of our model are: Dirichlet (hyper)parameters $\alpha_1, \ldots, \alpha_K$ for the generating distribution of the membership scores and aspect multinomial probabilities for words $\theta_{1k}(w) = p(w|k)$ and references $\theta_{2k}(r) = q(r|k)$, $k = 1, 2, \ldots, K$.

In the generative model, documents $d = (\{x_1^{(r_1)}\}, \{x_2^{(r_2)}\})$ are sampled according to the following sequence,

$$\lambda \sim \text{Dirichlet}(\alpha), \qquad [4]$$

$$x_1^{(r_1)} \sim \text{multinomial}(p_\lambda), \quad \text{where } p_\lambda = \sum_{k=1}^{K} \lambda_k \theta_{1k}, \qquad [5]$$

$$x_2^{(r_2)} \sim \text{multinomial}(q_\lambda), \quad \text{where } q_\lambda = \sum_{k=1}^{K} \lambda_k \theta_{2k}, \qquad [6]$$

where $\sum_w \theta_{1k}(w) = 1$ and $\sum_r \theta_{2k}(r) = 1$, $k = 1, \ldots, K$. Because distributions of words and references in a document are convex combinations of the distributions of the aspects, the aspects can be thought of as extreme or basis categories for a collection of documents. The sampling of words and references in the model

can be interpreted also as a latent classification process in which an aspect of origin is drawn first for each word and for each reference in a document, according to a multinomial distribution parameterized by the document-specific membership scores λ, and words and references then are generated from corresponding distributions of the aspects of origin (6). Rather than a mixture of K latent classes, the model can be thought of as a "simplicial mixture" (13) because the word and reference probabilities range over a simplex with corners θ_{1k} and θ_{2k}, respectively.

The likelihood function is thus

$$p(\boldsymbol{\theta}|d) = \int \mathrm{Dir}(\lambda|\alpha)\prod_w p_\lambda(w)^{n(w,d)} \prod_r q_\lambda(r)^{n(r,d)} d\lambda \qquad [7]$$

$$= \frac{\Gamma(\Sigma_i\, \alpha_i)}{\Pi_i\, \Gamma(\alpha_i)} \int \prod_{i=1}^{k} \lambda_i^{\alpha_i - 1} \prod_w p_\lambda(w)^{n(w,d)} \prod_r q_\lambda(r)^{n(r,d)} d\lambda, \qquad [8]$$

where integrals are over the $(K - 1)$ simplex.

It is important to note that the assumption of exchangeability among words and references (conditional independence given the membership scores) does not imply joint independence among the observed characteristics. Instead, the assumption of exchangeability means that dependencies among words and references can be explained fully by the membership scores of the documents. For an extended discussion on exchangeability in this context, see ref. 16.

Alternative Model for References

For the analysis of PNAS publications in the next section, we assume multinomial sampling of words and references. Although multinomial sampling is computationally convenient, it is not a realistic model of the way in which authors select references for the bibliography of an article. We briefly describe an example of more realistic generative assumptions for references.

Suppose an article focuses on a sufficiently narrow scientific area. In this case, the authors may have essentially perfect knowledge of the literature, and thus they would pay separate attention to each article in their pool of references as they consider whether to include it in the bibliography. Under these circumstances, given that the pool of references contains R articles, we assume that a document is represented as $d = (\{x_1^{(r_1)}\}, x_2, x_3, \ldots, x_{R+1})$, where $x_1^{(r_1)}$ is a word in the abstract, R is the number of references, and $x_2, \ldots, x_{R+1}$ are all references in the pool. Reference counts do not change: they are given by $n(r, d) = 1$ if the bibliography of d contains a reference to r and by $n(r, d) = 0$ if otherwise.

Then our model for generating documents would be to sample λ and $x_1^{(r_1)}$, according to Eqs. **4** and **5**, and sample x_j, $j = 2, \ldots, R + 1$, according to

$$x_j \sim \mathrm{Bernoulli}[q_\lambda(x_j)], \quad \text{where } q_\lambda(x_j) = \sum_{k=1}^{K} \lambda_k \theta_{jk}. \qquad [9]$$

The likelihood function based on this alternative model would not only take into account which documents contain which references, but it also would incorporate the information about which references documents do not contain.

Both the basic model for references and any alternatives still would need to reflect the time ordering on publications and include in the pool of possible references only those that have been published already, perhaps even with a short time lag. However, even such changes are unlikely to produce a "correct" model for citation practices.

Estimating the Model

The primary complication in using a mixed-membership model such as is shown in Eqs. **4–6**, in which the membership probabilities are random rather than fixed, is that the integral in Eq. **7** cannot be computed explicitly and therefore must be approximated. Two approximation schemes have been investigated recently for this problem and the associated problem of fitting the model. In the variational approach (12), the mixture terms $p_\lambda(w) = \sum_{k=1}^{K} \lambda_k \theta_{1k}(w)$ are bounded from below in a product form that leads to a tractable integral; the lower bound is then maximized. A related approach, called expectation–propagation (13), also approximates each mixture term in a product form but chooses the parameters of the factors by matching first and second moments. Either of these approximations to the integral (Eq. **7**) can be used in an approximate expectation–maximization (EM) algorithm to estimate the parameters of the models. It is shown in ref. 13 that expectation–propagation in general leads to better approximations than the simple variational method for mixed-membership models, although we obtained comparable results with both approaches on the PNAS collection. The results reported below use the variational approximation.

The PNAS Database

The National Academy of Sciences provided the database for the participants of the colloquium. We focused on a subset of all biological sciences articles in volumes 94–98 (Julian years 1997–2001) of PNAS, thereby ignoring articles published in the social and physical sciences unless they have official dual classifications with one classification in the biological sciences. The reason for this narrowing of focus is 2-fold. First, the major share of PNAS publications in recent years represents research developments in the biological sciences. Thus, of 13,008 articles published in volumes 94–98, 12,036 (92.53%) are in the biological sciences. The share of social and physical sciences articles in volumes 94–98 is a much more modest 7.47%. Second, we assume that a collection of articles is characterized by mixed membership in a number of internal categories, and social and physical sciences articles are unlikely to share the same internal categories with articles from the biological sciences. We also automatically ignore other types of PNAS publications such as corrections, commentaries, letters, and reviews, because these are not traditional research reports. Among the biological sciences articles in our database, 11 articles were not processed because they did not have an abstract, and 1 article was not processed because it did not contain any references.

PNAS is one of world's most cited multidisciplinary scientific journals. Historically, when submitting a research paper to PNAS, authors have to select a major category from physical, biological, or social sciences and a minor category from the list of topics. PNAS permits dual classifications between major categories and, in exceptional cases, within a major category. The lists of topics change over time to reflect changes in the National Academy of Sciences sections. PNAS, in its information for authors (revised in June 2002), states that it classifies publications in biological sciences according to 19 topics; the numbers of published articles and numbers of dual-classified articles in each topic are shown in Table 1.

The topic labels provide a classification structure for published materials, and most of the articles are members of only a single topic. For our mixed-membership model, we assume that there is a fixed number of extreme internal categories or aspects, each of which is characterized by multinomial distributions over words (in abstracts) and references (in bibliographies). Aspects are determined from contextual decompositions in such a way

Table 1. Biological sciences publications in PNAS volumes 94–98 by subtopic

Topic		*n*
1	Biochemistry	2,578 (33)
2	Medical sciences	1,547 (13)
3	Neurobiology	1,343 (9)
4	Cell biology	1,231 (10)
5	Genetics	980 (14)
6	Immunology	865 (9)
7	Biophysics	636 (40)
8	Evolution	510 (12)
9	Microbiology	498 (11)
10	Plant biology	488 (4)
11	Developmental biology	366 (2)
12	Physiology	340 (1)
13	Pharmacology	188 (2)
14	Ecology	133 (5)
15	Applied biological sciences	94 (6)
16	Psychology	88 (1)
17	Agricultural sciences	43 (2)
18	Population biology	43 (5)
19	Anthropology	10 (0)
	Total	11,981 (179)

The numbers of articles with dual classifications are given in parentheses.

that a multinomial distribution of words and references in each document is a convex combination of the corresponding distributions from the aspects. The convex combination for each article is based on proportions of the article's content coming from each category. These proportions, or membership scores, determine soft classifications of articles with respect to internal categories.

Results

Choosing a suitable value for the number of internal categories or aspects, K, in this type of setting is difficult. In our analyses, we focused largely on two versions of the model: one with 8 aspects and the other with 10. The set of parameters in our model is given by multinomial word and reference probabilities for each aspect and by the parameters of Dirichlet distribution, which is a generating distribution for membership scores. There are 39,616 unique words and 77,115 unique references in our data; hence, adding an aspect corresponds to having 39,615 + 77,114 + 1 = 116,730 additional parameters. Because of the large numbers of parameters involved, it is difficult to assess the extent to which the added pair of aspects actually improves the fit of the model to the data. On the basis of a set of preliminary comparisons, we found little to choose between them in fit and greater ease of interpretation for the eight-aspect model. Therefore, we report only the results of the eight-aspect model here.

To determine whether there are certain contexts that correspond to the aspects, we examine the most common words in the estimated multinomial distributions. In Table 2, we report the first 15 of the high-probability words for each aspect, filtering out so-called stop words, words that are generally common in English. An alternative way would be to discard the words from the "stop list" before fitting the model. If the distribution of stop words is not uniform across the internal categories, this alternative approach may potentially produce different results.

The following interpretations are based on examination of 50 high-probability words for each aspect. Note that enumeration of the aspects is arbitrary. The first aspect includes words such as Ca^{2+}, kinase, phosphorylation, receptor, and G (protein) channel, which pertain to cell signaling and intracellular signal transduction. It is likely that, in this aspect, signal transduction

Table 2. High-probability words for each aspect

Aspect 1	*P*	Aspect 2	*P*	Aspect 3	*P*	Aspect 4	*P*	Aspect 5	*P*	Aspect 6	*P*	Aspect 7	*P*	Aspect 8	*P*
Ca^{2+}	0.0062	species	0.0040	sequence	0.0024	development	0.0034	residues	0.0028	transcription	0.0060	IL	0.0046	increased	0.0027
channel	0.0047	sequence	0.0026	acid	0.0020	neurons	0.0034	enzyme	0.0023	nuclear	0.0036	tumor	0.0040	receptors	0.0023
membrane	0.0047	sequences	0.0024	plants	0.0018	brain	0.0029	active	0.0020	promoter	0.0031	activation	0.0036	G	0.0022
channels	0.0040	genetic	0.0024	cDNA	0.0017	mouse	0.0025	terminal	0.0019	transcriptional	0.0030	HIV	0.0032	*P*	0.0022
receptors	0.0028	genome	0.0022	mutant	0.0015	normal	0.0024	amino	0.0019	p53	0.0029	apoptosis	0.0031	insulin	0.0018
synaptic	0.0026	evolution	0.0020	single	0.0015	expressed	0.0021	RNA	0.0018	RNA	0.0027	kinase	0.0028	effects	0.0018
neurons	0.0022	among	0.0017	enzyme	0.0015	cortex	0.0019	structural	0.0018	kinase	0.0024	antigen	0.0026	increase	0.0018
G	0.0021	population	0.0016	plant	0.0014	embryonic	0.0017	state	0.0018	yeast	0.0024	virus	0.0025	acid	0.0018
calcium	0.0021	most	0.0016	identified	0.0013	adult	0.0017	folding	0.0017	function	0.0022	gamma	0.0021	effect	0.0016
activation	0.0020	chromosome	0.0015	amino	0.0013	neuronal	0.0016	sequence	0.0017	activation	0.0020	infection	0.0021	fold	0.0016
release	0.0020	selection	0.0015	expressed	0.0013	function	0.0016	form	0.0016	sequence	0.0018	immune	0.0020	reduced	0.0016
kinase	0.0019	populations	0.0014	mutants	0.0013	neural	0.0015	peptide	0.0016	terminal	0.0018	signaling	0.0018	treatment	0.0016
subunit	0.0019	three	0.0014	molecules	0.0012	early	0.0014	ATP	0.0015	cycle	0.0018	death	0.0017	glucose	0.0016
intracellular	0.0017	based	0.0013	based	0.0012	patients	0.0014	helix	0.0015	mutations	0.0017	activated	0.0017	mRNA	0.0015
acid	0.0016	variation	0.0013	kDa	0.0011	functional	0.0013	substrate	0.0015	factors	0.0017	*vivo*	0.0017	rats	0.0015

Table 3. High-probability references by aspect

Aspect 1

Author	Journal, Year	C
HAMILL OP	PFLUG ARCH EUR J PHY, 1981	72
LAEMMLI UK	Nature, 1970	322
HILLE B	IONIC CHANNELS EXCIT, 1992	58
BLISS TVP	NATURE, 1993	54
SUDHOF TC	NATURE, 1995	33
GRYNKIEWICZ G	J BIOL CHEM, 1985	31
SAMBROOK J	MOL CLONING LAB MANU, 1989	764
SHERRINGTON R	NATURE, 1995	33
ROTHMAN JE	NATURE, 1994	27
SIMONS K	NATURE, 1997	35
SOLLNER T	NATURE, 1993	25
ROTHMAN JE	SCIENCE, 1996	24
THINAKARAN G	NEURON, 1996	23
TOWBIN H	P NATL ACAD SCI USA, 1979	86
BERMAN DM	CELL, 1996	21

Aspect 2

Author	Journal, Year	C
SAITOU N	MOL BIOL EVOL, 1987	96
THOMPSON JD	NUCLEIC ACIDS RES, 1994	147
ALTSCHUL SF	NUCLEIC ACIDS RES, 1997	160
SAMBROOK J	MOL CLONING LAB MANU, 1989	764
ALTSCHUL SF	J MOL BIOL, 1990	253
FELSENSTEIN J	EVOLUTION, 1985	51
KISHINO H	J MOL EVOL, 1989	31
STRIMMER K	MOL BIOL EVOL, 1996	31
KIMURA M	J MOL EVOL, 1980	34
EISEN MB	P NATL ACAD SCI USA, 1998	60
SWOFFORD DL	PAUP PHYLOGENETIC AN, 1993	25
KIMURA M	NEUTRAL THEORY MOL E, 1983	28
KUMAR S	MEGA MOL EVOL UTIONAR, 1993	26
HASEGAWA M	J MOL EVOL, 1985	24
NEI M	MOL EVOLUTIONARY GEN, 1987	28

Aspect 3

Author	Journal, Year	C
SAMBROOK J	MOL CLONING LAB MANU, 1989	764
LAEMMLI UK	NATURE, 1970	322
ALTSCHUL SF	J MOL BIOL, 1990	253
BRADFORD MM	ANAL BIOCHEM, 1976	209
SANGER F	P NATL ACAD SCI USA, 1977	140
MILLER JH	EXPT MOL GENETICS, 1972	102
ALTSCHUL SF	NUCLEIC ACIDS RES, 1997	160
THOMPSON JD	NUCLEIC ACIDS RES, 1994	147
CHOMCZYNSKI P	ANAL BIOCHEM, 1987	206
HARLOW E	ANTIBODIES LAB MANUA, 1988	129
BLATTNER FR	SCIENCE, 1997	56
SCHENA M	SCIENCE, 1995	40
KYTE J	J MOL BIOL, 1982	51
MURASHIGE T	PHYSL PLANTARUM, 1962	33
TOWBIN H	P NATL ACAD SCI USA, 1979	86

Aspect 4

Author	Journal, Year	C
HOGAN B	MANIPULATING MOUSE E, 1994	68
CHOMCZYNSKI P	ANAL BIOCHEM, 1987	206
TALAIRACH J	COPLANAR STEREOTAXIC, 1988	60
PAXINOS G	RAT BRAIN STEREOTAXI, 1986	38
SAMBROOK J	MOL CLONING LAB MANU, 1989	764
NAGY A	P NATL ACAD SCI USA, 1993	39
MANSOUR SL	NATURE, 1988	37
BRAND AH	DEVELOPMENT, 1993	46
HOGAN B	MANIPULATING MOUSE E, 1986	32
TYBULEWICZ VLJ	CELL, 1991	46
KWONG KK	P NATL ACAD SCI USA, 1992	24
DUNLAP JC	CELL, 1999	19
LI E	CELL, 1992	35
ALTSCHUL SF	J MOL BIOL, 1990	253
EISEN MB	P NATL ACAD SCI USA, 1998	60

Aspect 5

Author	Journal, Year	C
KRAULIS PJ	J APPL CRYSTALLOGR, 1991	202
JONES TA	ACTA CRYSTALLOGR A, 1991	174
OTWINOWSKI Z	METHOD ENZYMOL, 1997	140
BRUNGER AT	ACTA CRYSTALLOGR D S, 1998	118
LASKOWSKI RA	J APPL CRYSTALLOGR, 1993	96
NICHOLLS A	PROTEINS, 1991	85
NAVAZA J	ACTA CRYSTALLOGR A, 1994	81
SAMBROOK J	MOL CLONING LAB MANU, 1989	764
LAEMMLI UK	NATURE, 1970	322
MERRITT EA	ACTA CRYSTALLOGR D, 1994	66
BRUNGER AT	NATURE, 1992	48
BRADFORD MM	ANAL BIOCHEM, 1976	209
MERRITT EA	METHOD ENZYMOL, 1997	41
WUTHRICH K	NMR PROTEINS NUCL AC, 1986	40
KABSCH W	BIOPOLYMERS, 1983	39

Aspect 6

Author	Journal, Year	C
SAMBROOK J	MOL CLONING LAB MANU, 1989	764
SIKORSKI RS	GENETICS, 1989	102
DIGNAM JD	NUCLEIC ACIDS RES, 1983	68
LEVINE AJ	CELL, 1997	57
ELDEIRY WS	CELL, 1993	54
HARLOW E	ANTIBODIES LAB MANUA, 1988	129
HARPER JW	CELL, 1993	50
FRIEDBERG EC	DNA REPAIR MUTAGENES, 1995	58
ALTSCHUL SF	J MOL BIOL 1990	253
OGRYZKO VV	CELL, 1996	41
WEINBERG RA	CELL, 1995	40
KAMEI Y	CELL, 1996	39
HOLLSTEIN M	SCIENCE, 1991	41
FIELDS S	NATURE, 1989	67
YANG XJ	NATURE, 1996	37

Aspect 7

Author	Journal, Year	C
DENG HK	NATURE, 1996	46
DRAGIC T	NATURE, 1996	45
DORANZ BJ	CELL, 1996	45
FENG Y	SCIENCE, 1996	43
ALKHATIB G	SCIENCE, 1996	43
COCCHI F	SCIENCE, 1995	41
CHOE H	CELL, 1996	41
THOMPSON CB	SCIENCE, 1995	38
ZOU H	CELL, 1997	40
DARNELL JE	SCIENCE, 1994	40
MUZIO M	CELL, 1996	35
LI P	CELL, 1997	36
XIA ZG	SCIENCE, 1995	38
BOLDIN MP	CELL, 1996	34
PEAR WS	P NATL ACAD SCI USA 1993	57

Aspect 8

Author	Journal, Year	C
CHOMCZYNSKI P	ANAL BIOCHEM, 1987	206
BRADFORD MM	ANAL BIOCHEM, 1976	209
LAEMMLI UK	NATURE, 1970	322
LOWRY OH	J BIOL CHEM, 1951	73
ZHANG Y	NATURE, 1994	31
KUIPER GGJM	P NATL ACAD SCI USA, 1996	27
SAMBROOK J	MOL CLON LAB MANU, 1989	764
MONCADA S	PHARMACOL REV, 1991	25
PELLEYMOUNTER MA	SCIENCE, 1995	23
CAMPFIELD LA	SCIENCE, 1995	23
KUIPER GGJM	ENDOCRINOLOGY, 1997	22
HALAAS JL	SCIENCE, 1995	21
BLIGH EG	CAN J BIOCH PHYSL, 1959	45
BROWN MS	CELL, 1997	28
ZHANG SH	SCIENCE 1992	18

For each aspect, the top references are shown in order of decreasing probability, according to the model. The count of each reference in the PNAS collection is shown in the right column (C).

is considered as applied to neuron signaling as indicated by the words synaptic, neurons, voltage. It is interesting that Ca^{2+} in the first aspect is the highest-probability contextual word over all the aspects. Frequent words for the second aspect indicate that its context is related to molecular evolution that deals with natural selection on the population and intraspecies level and mechanisms of acquiring genetic traits. Words in aspect 3 pertain mostly to the plant molecular biology area. High-probability words in aspect 4 relate to studies of neuronal responses in mice and humans, which identify this aspect as related to developmental biology and neurobiology. Aspect 5 contains words that can be associated with biochemistry and molecular biology. Words in aspect 6 point to genetics and molecular biology. Frequent words for aspect 7 contain such terms as immune, IL (or interleukin), antigen, (IFN) gamma, and MHC class II, which point to a relatively new area in immunology, namely, tumor immunology. The presence of such words as HIV and virus in aspect 7 indicates a more general immunology content. For aspect 8, words such as increase or reduced, treatment, effect, fold, and P (assuming it stands for P value) correspond to general reporting of experimental results, likely in the area of endocrinology.

As for words, multinomial distributions are estimated for the references that are present in our collection. For estimation, we

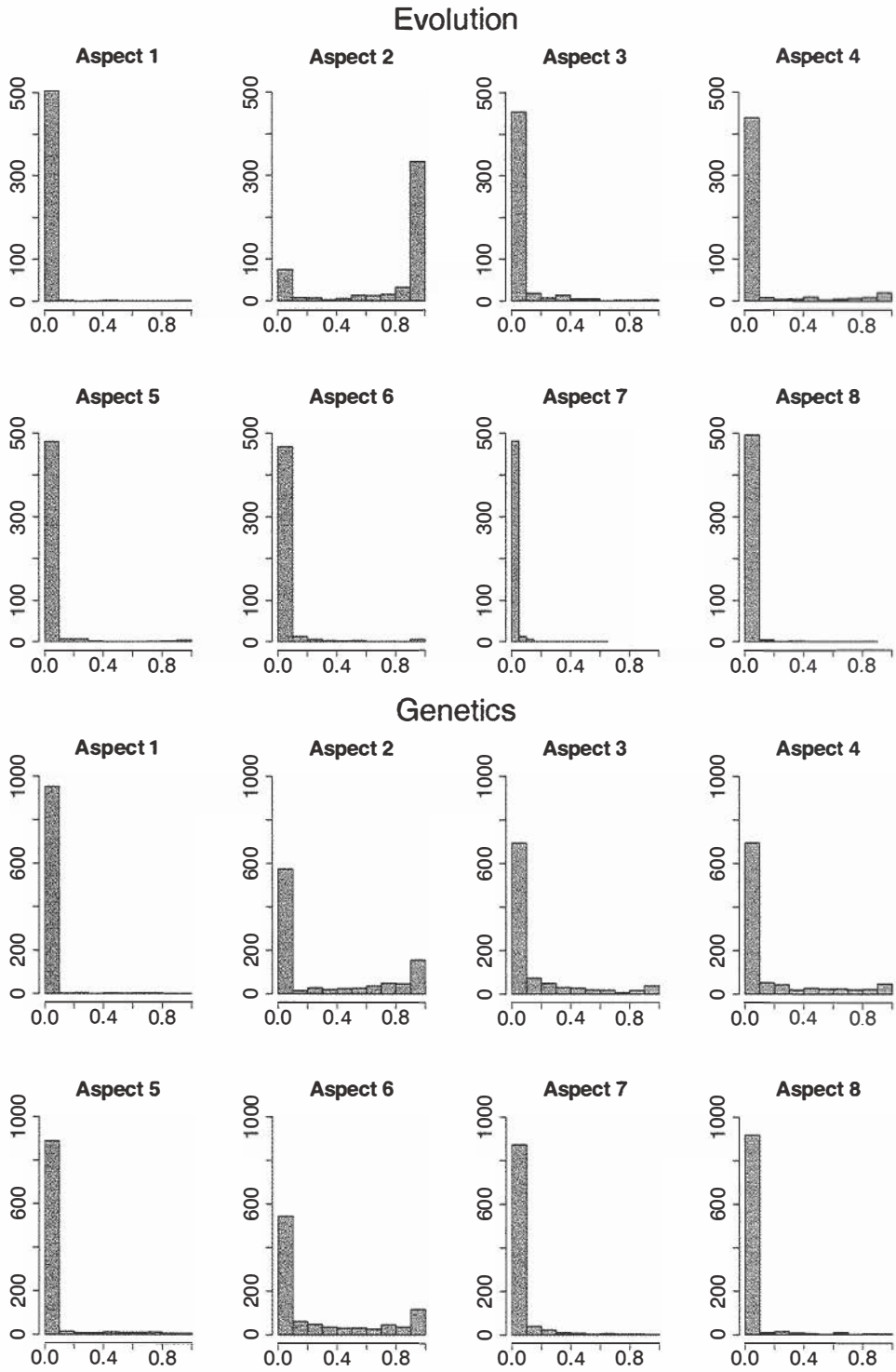
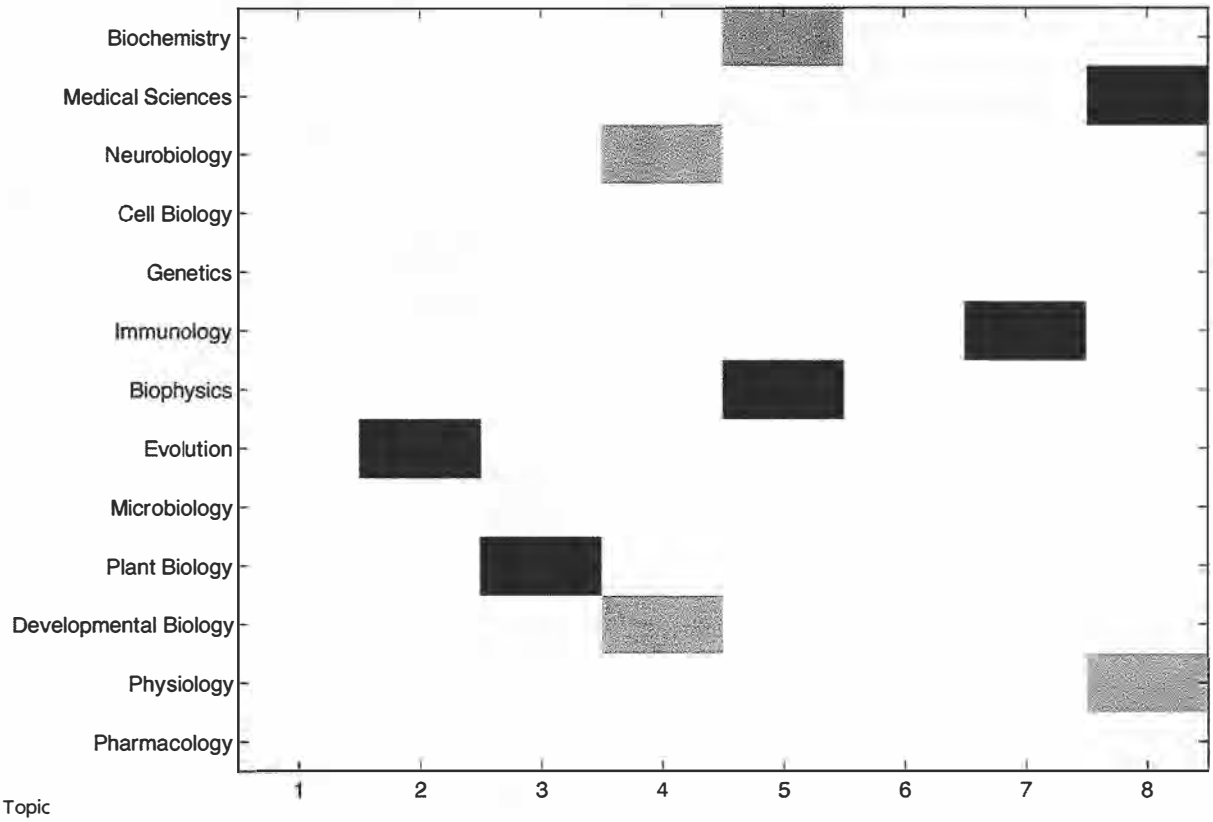


Fig. 1. Distributions by aspect of the posterior means of membership scores for articles published in evolution and genetics.

only need unique indicators for each referenced article. After the model is fitted, attributes of high-probability references for each aspect provide additional information about its contextual interpretation. Table 3 provides attributes of 15 high-probability references for each aspect that were available in the database together with PNAS citation counts (number of times cited by PNAS articles in the database). Notice that, because the model draws from the contextual decomposition, having a high citation count is not necessary for having high aspect probability. In Table 3, high-probability references for aspect 1 are dominated by publications in *Nature*; references in aspect 7 are mostly *Nature*, *Cell*, and *Science* publications from the mid-1990s.

Examining titles of the references (see Table 5, which is published as supporting information on the PNAS web site, www.pnas.org), we see that manuals, textbooks, and references to methodology articles seem to be prominent for many aspects. Thus, among the first 15 high-probability references, all 15 from aspect 3 and more than half from aspect 4 are of this method-

Table 4. Mean decompositions of aspect membership scores (*Lower*), together with a graphical representation of this table (*Upper*)



Topic	1	2	3	4	5	6	7	8
Biochemistry	0.0469	0.0347	0.1810	0.0178	0.3838	0.2057	0.0477	0.0823
Medical sciences	0.0244	0.0502	0.0938	0.1274	0.0181	0.1075	0.3286	0.2500
Neurobiology	0.2875	0.0398	0.0722	0.3768	0.0196	0.0296	0.0441	0.1304
Cell biology	0.1691	0.0165	0.1420	0.0684	0.1097	0.2423	0.1637	0.0884
Genetics	0.0141	0.3056	0.1422	0.1532	0.0487	0.2621	0.0395	0.0347
Immunology	0.0127	0.0593	0.1003	0.0413	0.0422	0.0915	0.6244	0.0283
Biophysics	0.0507	0.0295	0.2398	0.0162	0.5496	0.0542	0.0176	0.0423
Evolution	0.0042	0.7679	0.0465	0.0913	0.0289	0.0378	0.0101	0.0133
Microbiology	0.0158	0.1725	0.3431	0.0335	0.0647	0.1174	0.1870	0.0661
Plant biology	0.1333	0.0983	0.4400	0.0360	0.0462	0.0954	0.0166	0.1344
Developmental biology	0.0475	0.0288	0.1071	0.3729	0.0274	0.2558	0.0974	0.0631
Physiology	0.3179	0.0275	0.0712	0.1123	0.0258	0.0116	0.0595	0.3743
Pharmacology	0.2883	0.0161	0.0772	0.1965	0.0299	0.0349	0.0537	0.3033

For clarity, the six lowest-frequency topics, which make up 3.4% of the biological sciences articles, are not shown.

ological type. In contrast, most high-probability references for aspect 7 are those that report new findings. Titles of the references indicate neurobiology content for aspect 1, molecular evolution for aspect 2, and plant molecular biology for aspect 3, which is in agreement with our conclusions based on high-probability words. For other aspects, titles of high-probability references help us refine the aspects. Thus, aspect 4 mostly pertains to the study of brain development, in particular, via genetic manipulation of mouse embryo. Aspect 5, identified as biochemistry and molecular biology by the words, can be described as protein structural biology by the references. Aspect 6 may be labeled in a more detailed way as "DNA repair, mutagenesis, and cell cycle." The references for aspects 7 and 8 shift their focuses more toward HIV infection and studies of molecular mechanisms of obesity.

Among frequent references for the eight aspects, there are seven PNAS articles that share a special feature: they were all either coauthored or contributed by a distinguished member of the National Academy of Sciences. In fact, one article was coauthored by a Nobel prize winner, and two were contributed by other Nobelists. Although these articles do not have the highest counts in the database, they are notable for various reasons; e.g., one is on clustering and gene expression (2), and it is also one of the two highly cited PNAS articles on clustering that we mentioned in the Introduction. These seven articles may not necessarily be off-beat, but they may be among those that fulfill MacLane's petition regarding the special nature of PNAS.

From our analysis of high-probability words, it is difficult to determine whether the majority of aspects correspond to a single topic from the official classifications in PNAS biological science publications. To investigate whether there is a correspondence between the estimated aspects and the given topics, we examine aspect loadings (means of posterior membership scores) for each article. Given estimated parameters of the model, the distribu-

tion of each article's loadings can be obtained by means of Bayes' theorem. The variational and expectation–propagation procedures provide Dirichlet approximations to the posterior distribution $p(\boldsymbol{\lambda}|d, \boldsymbol{\theta})$ for each document d. We use the mean of this Dirichlet as an estimate of the weight of the document on each aspect. Histograms of these loadings are provided in Fig. 1 for articles in evolution and genetics. Relatively high histogram bars near zero correspond to the majority of articles having small posterior membership scores for the given aspect. Among the articles published in genetics, some can be considered as full members in aspects 2, 3, 4, and 6, but many have mixed membership in these and other aspects. Articles published in evolution, on the other hand, show a somewhat different behavior: the majority of these articles comes fully from aspect 2.

The sparsity of the loadings can be gauged also by the parameters of the Dirichlet distribution, which are estimated as $\alpha_1 = 0.0195$, $\alpha_2 = 0.0203$, $\alpha_3 = 0.0569$, $\alpha_4 = 0.0346$, $\alpha_5 = 0.0317$, $\alpha_6 = 0.0363$, $\alpha_7 = 0.0411$, and $\alpha_8 = 0.0255$. The estimated Dirichlet, which is the generative distribution of membership scores, is "bathtub-shaped" on the simplex; as a result, articles tend to have relatively high membership scores in only a few aspects.

To summarize the aspect distributions for each topic, we provide mean loadings and the graphical representation of these values in Table 4 *Upper*. Larger values correspond to darker colors, and the values below some threshold are not shown (white) for clarity. As an example, the mean loading of 0.2883 for pharmacology in the first aspect is the average of the posterior means of the membership scores for this aspect over all pharmacology publications in the database. Note that this percentage is based on the assumption of mixed membership and can be interpreted as indicating that 29% of the words in pharmacology articles originate from aspect 1, according to our model.

Examining the rows of Table 4, we see that most subtopics in biological sciences have major components from more than one aspect (extreme or basis category). Examining the columns, we can gain additional insights in interpretation of the extreme categories. Aspect 8, for example, is the aspect of origin for a combined 37% of physiology, 30% of pharmacology, and 25% of medical sciences articles, according to the mixed-membership model. The most prominent subtopic is evolution; it has the greatest influence in defining an extremal category, aspect 2. This is consistent with a special place that evolution holds among the biological sciences by standing apart both conceptually and methodologically.

Finally, we compare the loadings (posterior means of the membership scores) of dual-classified articles to those that are singly classified. We consider two articles as similar if their loadings are equal for the first significant digit for all aspects. One might interpret singly classified articles that are similar to dual-classified as articles that should have had dual classification but did not. We find that, for 11% of the singly classified articles, there is at least one similar dual-classified article. For example, three biophysics dual-classified articles with loadings 0.9 for the second and 0.1 for the third aspect turned out to be similar to 86 singly classified articles from biophysics, biochemistry, cell biology, developmental biology, evolution, genetics, immunology, medical sciences, and microbiology.

Concluding Remarks

We have presented results from fitting a mixed-membership model to PNAS biological sciences publications, from 1997 to 2001, providing an implicit semantic decomposition of words and references in the articles. The model allows us to identify extreme internal categories of publications and to provide soft classifications of articles into these categories. Our results show that the traditional discipline classifications correspond to a mixed distribution over the internal categories. Our analyses and modeling were intended to capture a high-level description of a subset of PNAS articles.

In an often-quoted statement, Box remarked: "all models are wrong" (17). In our case, the assumption of a bag of words and references in the mixed-membership model clearly oversimplifies reality; the model does not account for the general structure of the language, nor does it capture the compositional structure of bibliographies. Many interesting extensions of the basic model we have explored are possible, from hierarchical models of topics to more detailed models of citations and dynamic models of the evolution of scientific fields over time. Nevertheless, as Box notes, even wrong models may be useful. Our results indicate that mixed-membership models can be useful for analyzing the implicit structure of scientific publications.

We thank Dr. Anna Lokshin (University of Pittsburgh, Pittsburgh) for help with interpreting model results from a biologist's perspective. E.E. was supported by National Institutes of Health Grants 1 R01 AG023141-01 and R01 CA94212-01; S.F. was supported by National Institutes of Health Grant 1 R01 AG023141-01. J.L. was supported by National Science Foundation Grant CCR-0122581 and Advanced Research and Development Activity Contract MDA904-00-C-2106.

1. MacLane, S. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 5983–5985.
2. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 14863–14868.
3. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander E. S. & Golub, T. R. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 2907–2912.
4. Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. & Feldman, M. W. (2002) *Science* **298,** 2381–2385.
5. Woodbury, M. A., Clive, J. & Garson, A. (1978) *Comput. Biomed. Res.* **11,** 277–298.
6. Erosheva, E. A. (2002) Ph.D. thesis (Carnegie Mellon University, Pittsburgh).
7. Griffiths, T. L. & Steyvers, M. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5228–5235.
8. Manton, K. G., Woodbury, M. A. & Tolley, H. D. (1994) *Statistical Applications Using Fuzzy Sets* (Wiley Interscience, New York), p. 312.
9. Potthoff, R. G., Manton, K. G., Woodbury, M. A. & Tolley, H. D. (2000) *J. Classification* **17,** 315–353.
10. Pritchard, J. K., Stephens, M. & Donnelly, P. (2000) *Genetics* **155,** 945–959.
11. Hofmann, T. (2001) *Machine Learn.* **42,** 177–196.
12. Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003) *J. Machine Learn. Res.* **3,** 993–1002.
13. Minka, T. P. & Lafferty, J. (2002) *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)* (Morgan Kaufmann, San Francisco), pp. 352–359.
14. Cohn, D. & Hofmann, T. (2001) *Neural Information Processing Systems 13* (MIT Press, Cambridge, MA).
15. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M. & Jordan, M. I. (2003) *J. Machine Learn. Res.* **3,** 1107–1135.
16. Blei, D. M., Jordan, M. I. & Ng, A. Y. (2003) in *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, eds. Bernardo, J. M., Bayarri, M. J., Dawid, A. P., Berger, J. O., Heckerman, D., Smith, A. F. M. & West, M. (Oxford Univ. Press, Oxford), pp. 25–44.
17. Box, G. E. P. (1979) in *Robustness in Statistics*, eds. Launer, R. L. & Wilkinson, G. G. (Academic, New York), p. 202.

Finding scientific topics

Thomas L. Griffiths*†‡ and Mark Steyvers§

*Department of Psychology, Stanford University, Stanford, CA 94305; †Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139-4307; and §Department of Cognitive Sciences, University of California, Irvine, CA 92697

A first step in identifying the content of a document is determining which topics that document addresses. We describe a generative model for documents, introduced by Blei, Ng, and Jordan [Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003) *J. Machine Learn. Res.* 3, 993-1022], in which each document is generated by choosing a distribution over topics and then choosing each word in the document from a topic selected according to this distribution. We then present a Markov chain Monte Carlo algorithm for inference in this model. We use this algorithm to analyze abstracts from PNAS by using Bayesian model selection to establish the number of topics. We show that the extracted topics capture meaningful structure in the data, consistent with the class designations provided by the authors of the articles, and outline further applications of this analysis, including identifying "hot topics" by examining temporal dynamics and tagging abstracts to illustrate semantic content.

When scientists decide to write a paper, one of the first things they do is identify an interesting subset of the many possible topics of scientific investigation. The topics addressed by a paper are also one of the first pieces of information a person tries to extract when reading a scientific abstract. Scientific experts know which topics are pursued in their field, and this information plays a role in their assessments of whether papers are relevant to their interests, which research areas are rising or falling in popularity, and how papers relate to one another. Here, we present a statistical method for automatically extracting a representation of documents that provides a first-order approximation to the kind of knowledge available to domain experts. Our method discovers a set of topics expressed by documents, providing quantitative measures that can be used to identify the content of those documents, track changes in content over time, and express the similarity between documents. We use our method to discover the topics covered by papers in PNAS in a purely unsupervised fashion and illustrate how these topics can be used to gain insight into some of the structure of science.

The statistical model we use in our analysis is a generative model for documents; it reduces the complex process of producing a scientific paper to a small number of simple probabilistic steps and thus specifies a probability distribution over all possible documents. Generative models can be used to postulate complex latent structures responsible for a set of observations, making it possible to use statistical inference to recover this structure. This kind of approach is particularly useful with text, where the observed data (the words) are explicitly intended to communicate a latent structure (their meaning). The particular generative model we use, called Latent Dirichlet Allocation, was introduced in ref. 1. This generative model postulates a latent structure consisting of a set of topics; each document is produced by choosing a distribution over topics, and then generating each word at random from a topic chosen by using this distribution.

The plan of this article is as follows. In the next section, we describe Latent Dirichlet Allocation and present a Markov chain Monte Carlo algorithm for inference in this model, illustrating the operation of our algorithm on a small dataset. We then apply our algorithm to a corpus consisting of abstracts from PNAS from 1991 to 2001, determining the number of topics needed to account for the information contained in this corpus and extracting a set of topics. We use these topics to illustrate the relationships between different scientific disciplines, assessing trends and "hot topics" by analyzing topic dynamics and using the assignments of words to topics to highlight the semantic content of documents.

Documents, Topics, and Statistical Inference

A scientific paper can deal with multiple topics, and the words that appear in that paper reflect the particular set of topics it addresses. In statistical natural language processing, one common way of modeling the contributions of different topics to a document is to treat each topic as a probability distribution over words, viewing a document as a probabilistic mixture of these topics (1–6). If we have T topics, we can write the probability of the ith word in a given document as

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j), \qquad [1]$$

where z_i is a latent variable indicating the topic from which the ith word was drawn and $P(w_i|z_i = j)$ is the probability of the word w_i under the jth topic. $P(z_i = j)$ gives the probability of choosing a word from topics j in the current document, which will vary across different documents.

Intuitively, $P(w|z)$ indicates which words are important to a topic, whereas $P(z)$ is the prevalence of those topics within a document. For example, in a journal that published only articles in mathematics or neuroscience, we could express the probability distribution over words with two topics, one relating to mathematics and the other relating to neuroscience. The content of the topics would be reflected in $P(w|z)$; the "mathematics" topic would give high probability to words like theory, space, or problem, whereas the "neuroscience" topic would give high probability to words like synaptic, neurons, and hippocampal. Whether a particular document concerns neuroscience, mathematics, or computational neuroscience would depend on its distribution over topics, $P(z)$, which determines how these topics are mixed together in forming documents. The fact that multiple topics can be responsible for the words occurring in a single document discriminates this model from a standard Bayesian classifier, in which it is assumed that all the words in the document come from a single class. The "soft classification" provided by this model, in which each document is characterized in terms of the contributions of multiple topics, has applications in many domains other than text (7).

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

‡To whom correspondence should be addressed. E-mail: gruffydd@psych.stanford.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0307752101

Viewing documents as mixtures of probabilistic topics makes it possible to formulate the problem of discovering the set of topics that are used in a collection of documents. Given D documents containing T topics expressed over W unique words, we can represent $P(w|z)$ with a set of T multinomial distributions ϕ over the W words, such that $P(w|z = j) = \phi_w^{(j)}$, and $P(z)$ with a set of D multinomial distributions θ over the T topics, such that for a word in document d, $P(z = j) = \theta_j^{(d)}$. To discover the set of topics used in a corpus $\mathbf{w} = \{w_1, w_2, \ldots, w_n\}$, where each w_i belongs to some document d_i, we want to obtain an estimate of ϕ that gives high probability to the words that appear in the corpus. One strategy for obtaining such an estimate is to simply attempt to maximize $P(\mathbf{w}|\phi, \theta)$, following from Eq. **1** directly by using the Expectation-Maximization (8) algorithm to find maximum likelihood estimates of ϕ and θ (2, 3). However, this approach is susceptible to problems involving local maxima and is slow to converge (1, 2), encouraging the development of models that make assumptions about the source of θ.

Latent Dirichlet Allocation (1) is one such model, combining Eq. **1** with a prior probability distribution on θ to provide a complete generative model for documents. This generative model specifies a simple probabilistic procedure by which new documents can be produced given just a set of topics ϕ, allowing ϕ to be estimated without requiring the estimation of θ. In Latent Dirichlet Allocation, documents are generated by first picking a distribution over topics θ from a Dirichlet distribution, which determines $P(z)$ for words in that document. The words in the document are then generated by picking a topic j from this distribution and then picking a word from that topic according to $P(w|z = j)$, which is determined by a fixed $\phi^{(j)}$. The estimation problem becomes one of maximizing $P(\mathbf{w}|\phi, \alpha) = \int P(\mathbf{w}|\phi,\theta)P(\theta|\alpha)d\theta$, where $P(\theta)$ is a Dirichlet (α) distribution. The integral in this expression is intractable, and ϕ is thus usually estimated by using sophisticated approximations, either variational Bayes (1) or expectation propagation (9).

Using Gibbs Sampling to Discover Topics

Our strategy for discovering topics differs from previous approaches in not explicitly representing ϕ or θ as parameters to be estimated, but instead considering the posterior distribution over the assignments of words to topics, $P(\mathbf{z}|\mathbf{w})$. We then obtain estimates of ϕ and θ by examining this posterior distribution. Evaluating $P(\mathbf{z}|\mathbf{w})$ requires solving a problem that has been studied in detail in Bayesian statistics and statistical physics, computing a probability distribution over a large discrete state space. We address this problem by using a Monte Carlo procedure, resulting in an algorithm that is easy to implement, requires little memory, and is competitive in speed and performance with existing algorithms.

We use the probability model for Latent Dirichlet Allocation, with the addition of a Dirichlet prior on ϕ. The complete probability model is thus

$$\begin{aligned} w_i|z_i, \phi^{(z_i)} &\sim \text{Discrete}(\phi^{(z_i)}) \\ \phi &\sim \text{Dirichlet}(\beta) \\ z_i|\theta^{(d_i)} &\sim \text{Discrete}(\theta^{(d_i)}) \\ \theta &\sim \text{Dirichlet}(\alpha) \end{aligned}$$

Here, α and β are hyperparameters, specifying the nature of the priors on θ and ϕ. Although these hyperparameters could be vector-valued as in refs. 1 and 9, for the purposes of this article we assume symmetric Dirichlet priors, with α and β each having a single value. These priors are conjugate to the multinomial distributions ϕ and θ, allowing us to compute the joint distribution $P(\mathbf{w}, \mathbf{z})$ by integrating out ϕ and θ. Because $P(\mathbf{w}, \mathbf{z}) = P(\mathbf{w}|\mathbf{z})P(\mathbf{z})$ and ϕ and θ only appear in the first and second terms, respectively, we can perform these integrals separately. Integrating out ϕ gives the first term

$$P(\mathbf{w}|\mathbf{z}) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{j=1}^{T} \frac{\Pi_w \Gamma(n_j^{(w)} + \beta)}{\Gamma(n_j^{(\cdot)} + W\beta)}, \quad [2]$$

in which $n_j^{(w)}$ is the number of times word w has been assigned to topic j in the vector of assignments $\mathbf{z}$, and $\Gamma(\cdot)$ is the standard gamma function. The second term results from integrating out θ, to give

$$P(\mathbf{z}) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D \prod_{d=1}^{D} \frac{\Pi_j \Gamma(n_j^{(d)} + \alpha)}{\Gamma(n_\cdot^{(d)} + T\alpha)}, \quad [3]$$

where $n_j^{(d)}$ is the number of times a word from document d has been assigned to topic j. Our goal is then to evaluate the posterior distribution.

$$P(\mathbf{z}|\mathbf{w}) = \frac{P(\mathbf{w}, \mathbf{z})}{\sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z})}. \quad [4]$$

Unfortunately, this distribution cannot be computed directly, because the sum in the denominator does not factorize and involves T^n terms, where n is the total number of word instances in the corpus.

Computing $P(\mathbf{z}|\mathbf{w})$ involves evaluating a probability distribution on a large discrete state space, a problem that arises often in statistical physics. Our setting is similar, in particular, to the Potts model (e.g., ref. 10), with an ensemble of discrete variables $\mathbf{z}$, each of which can take on values in $\{1, 2, \ldots, T\}$, and an energy function given by $H(\mathbf{z}) \propto -\log P(\mathbf{w}, \mathbf{z}) = -\log P(\mathbf{w}|\mathbf{z}) - \log P(\mathbf{z})$. Unlike the Potts model, in which the energy function is usually defined in terms of local interactions on a lattice, here the contribution of each z_i depends on all $\mathbf{z}_{-i}$ values through the counts $n_j^{(w)}$ and $n_j^{(d)}$. Intuitively, this energy function favors ensembles of assignments $\mathbf{z}$ that form a good compromise between having few topics per document and having few words per topic, with the terms of this compromise being set by the hyperparameters α and β. The fundamental computational problems raised by this model remain the same as those of the Potts model: We can evaluate $H(\mathbf{z})$ for any configuration $\mathbf{z}$, but the state space is too large to enumerate, and we cannot compute the partition function that converts this into a probability distribution (in our case, the denominator of Eq. **4**). Consequently, we apply a method that physicists and statisticians have developed for dealing with these problems, sampling from the target distribution by using Markov chain Monte Carlo.

In Markov chain Monte Carlo, a Markov chain is constructed to converge to the target distribution, and samples are then taken from that Markov chain (see refs. 10–12). Each state of the chain is an assignment of values to the variables being sampled, in this case $\mathbf{z}$, and transitions between states follow a simple rule. We use Gibbs sampling (13), known as the heat bath algorithm in statistical physics (10), where the next state is reached by sequentially sampling all variables from their distribution when conditioned on the current values of all other variables and the data. To apply this algorithm we need the full conditional distribution $P(z_i|\mathbf{z}_{-i}, \mathbf{w})$. This distribution can be obtained by a probabilistic argument or by cancellation of terms in Eqs. **2** and **3**, yielding

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}, \quad [5]$$

where $n_{-i}^{(\cdot)}$ is a count that does not include the current assignment of z_i. This result is quite intuitive; the first ratio expresses the probability of w_i under topic j, and the second ratio expresses the probability of topic j in document d_i. Critically, these counts are

the only information necessary for computing the full conditional distribution, allowing the algorithm to be implemented efficiently by caching the relatively small set of nonzero counts.

Having obtained the full conditional distribution, the Monte Carlo algorithm is then straightforward. The z_i variables are initialized to values in $\{1, 2, \ldots, T\}$, determining the initial state of the Markov chain. We do this with an on-line version of the Gibbs sampler, using Eq. **5** to assign words to topics, but with counts that are computed from the subset of the words seen so far rather than the full data. The chain is then run for a number of iterations, each time finding a new state by sampling each z_i from the distribution specified by Eq. **5**. Because the only information needed to apply Eq. **5** is the number of times a word is assigned to a topic and the number of times a topic occurs in a document, the algorithm can be run with minimal memory requirements by caching the sparse set of nonzero counts and updating them whenever a word is reassigned. After enough iterations for the chain to approach the target distribution, the current values of the z_i variables are recorded. Subsequent samples are taken after an appropriate lag to ensure that their autocorrelation is low (10, 11).

With a set of samples from the posterior distribution $P(\mathbf{z}|\mathbf{w})$, statistics that are independent of the content of individual topics can be computed by integrating across the full set of samples. For any single sample we can estimate ϕ and θ from the value $\mathbf{z}$ by

$$\hat{\phi}_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + W\beta} \quad [6]$$

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_\cdot^{(d)} + T\alpha}. \quad [7]$$

These values correspond to the predictive distributions over new words w and new topics z conditioned on $\mathbf{w}$ and $\mathbf{z}$.[¶]

A Graphical Example

To illustrate the operation of the algorithm and to show that it runs in time comparable with existing methods of estimating ϕ, we generated a small dataset in which the output of the algorithm can be shown graphically. The dataset consisted of a set of 2,000 images, each containing 25 pixels in a 5×5 grid. The intensity of any pixel is specified by an integer value between zero and infinity. This dataset is of exactly the same form as a word-document cooccurrence matrix constructed from a database of documents, with each image being a document, with each pixel being a word, and with the intensity of a pixel being its frequency. The images were generated by defining a set of 10 topics corresponding to horizontal and vertical bars, as shown in Fig. 1*a*, then sampling a multinomial distribution θ for each image from a Dirichlet distribution with $\alpha = 1$, and sampling 100 pixels (words) according to Eq. **1**. A subset of the images generated in this fashion are shown in Fig. 1*b*. Although some images show evidence of many samples from a single topic, it is difficult to discern the underlying structure of most images.

We applied our Gibbs sampling algorithm to this dataset, together with the two algorithms that have previously been used for inference in Latent Dirichlet Allocation: variational Bayes (1) and expectation propagation (9). (The implementations of variational Bayes and expectation propagation were provided by Tom Minka and are available at www.stat.cmu.edu/~minka/papers/aspect.html.) We divided the dataset into 1,000 training images and 1,000 test images and ran each algorithm four times, using the same initial conditions for all three algorithms on a given run. These initial conditions were found by an online application of Gibbs sampling, as mentioned above. Variational Bayes and expectation propagation were run until convergence, and Gibbs sampling was run for 1,000 iterations. All three algorithms used a fixed Dirichlet prior on θ, with $\alpha = 1$. We tracked the number of floating point operations per iteration for each algorithm and computed the test set perplexity for the estimates of ϕ provided by the algorithms at several points. Perplexity is a standard measure of performance for statistical models of natural language (14) and is defined as $\exp\{-\log P(\mathbf{w}_{\text{test}}|\phi)/n_{\text{test}}\}$, where $\mathbf{w}_{\text{test}}$ and n_{test} indicate the identities and number of words in the test set, respectively. Perplexity indicates the uncertainty in predicting a single word; lower values are better, and chance performance results in a perplexity equal to the size of the vocabulary, which is 25 in this case. The perplexity for all three models was evaluated by using importance sampling as in ref. 9, and the estimates of ϕ used for evaluating Gibbs sampling were each obtained from a single sample as in Eq. **6**. The results of these computations are shown in Fig. 1*c*. All three algorithms are able to recover the underlying topics, and Gibbs sampling does so more rapidly than either variational Bayes or expectation propagation. A graphical illustration of the operation of the Gibbs sampler is shown in Fig. 2. The log-likelihood stabilizes quickly, in a fashion consistent across multiple runs, and the topics expressed in the dataset slowly emerge as appropriate assignments of words to topics are discovered.

These results show that Gibbs sampling can be competitive in speed with existing algorithms, although further tests with larger datasets involving real text are necessary to evaluate the strengths and weaknesses of the different algorithms. The effects of including the Dirichlet (β) prior in the model and the use of methods for estimating the hyperparameters α and β need to be assessed as part of this comparison. A variational algorithm for

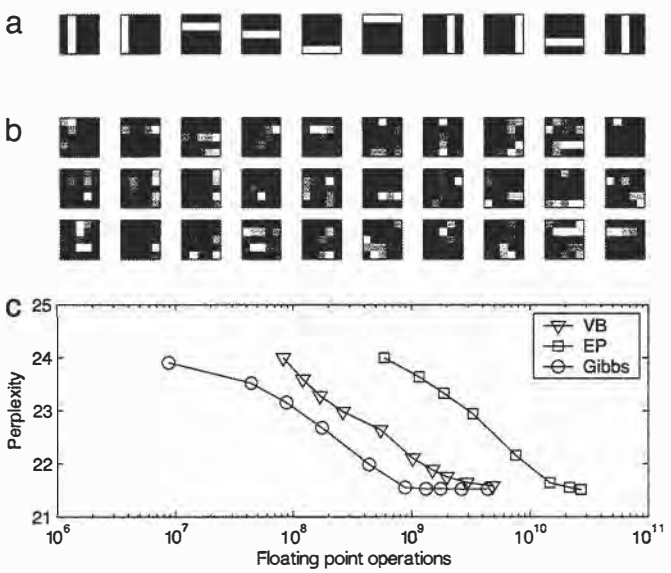


Fig. 1. (a) Graphical representation of 10 topics, combined to produce "documents" like those shown in *b*, where each image is the result of 100 samples from a unique mixture of these topics. (*c*) Performance of three algorithms on this dataset: variational Bayes (VB), expectation propagation (EP), and Gibbs sampling. Lower perplexity indicates better performance, with chance being a perplexity of 25. Estimates of the standard errors are smaller than the plot symbols, which mark 1, 5, 10, 20, 50, 100, 150, 200, 300, and 500 iterations.

[¶]These estimates cannot be combined across samples for any analysis that relies on the content of specific topics. This issue arises because of a lack of identifiability. Because mixtures of topics are used to form documents, the probability distribution over words implied by the model is unaffected by permutations of the indices of the topics. Consequently, no correspondence is needed between individual topics across samples; just because two topics have index j in two samples is no reason to expect that similar words were assigned to those topics in those samples. However, statistics insensitive to permutation of the underlying topics can be computed by aggregating across samples.

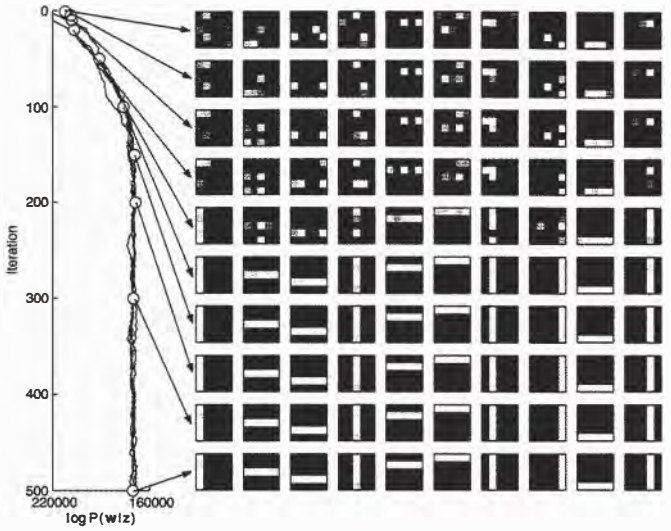


Fig. 2. Results of running the Gibbs sampling algorithm. The log-likelihood, shown on the left, stabilizes after a few hundred iterations. Traces of the log-likelihood are shown for all four runs, illustrating the consistency in values across runs. Each row of images on the right shows the estimates of the topics after a certain number of iterations within a single run, matching the points indicated on the left. These points correspond to 1, 2, 5, 10, 20, 50, 100, 150, 200, 300, and 500 iterations. The topics expressed in the data gradually emerge as the Markov chain approaches the posterior distribution.

this "smoothed" model is described in ref. 1, which may be more similar to the Gibbs sampling algorithm described here. Ultimately, these different approaches are complementary rather than competitive, providing different means of performing approximate inference that can be selected according to the demands of the problem.

Model Selection

The statistical model we have described is conditioned on three parameters, which we have suppressed in the equations above: the Dirichlet hyperparameters α and β and the number of topics T. Our algorithm is easily extended to allow α, β, and $\mathbf{z}$ to be sampled, but this extension can slow the convergence of the Markov chain. Our strategy in this article is to fix α and β and explore the consequences of varying T. The choice of α and β can have important implications for the results produced by the model. In particular, increasing β can be expected to decrease the number of topics used to describe a dataset, because it reduces the impact of sparsity in Eq. **2**. The value of β thus affects the granularity of the model: a corpus of documents can be sensibly factorized into a set of topics at several different scales, and the particular scale assessed by the model will be set by β. With scientific documents, a large value of β would lead the model to find a relatively small number of topics, perhaps at the level of scientific disciplines, whereas smaller values of β will produce more topics that address specific areas of research.

Given values of α and β, the problem of choosing the appropriate value for T is a problem of model selection, which we address by using a standard method from Bayesian statistics (15). For a Bayesian statistician faced with a choice between a set of statistical models, the natural response is to compute the posterior probability of that set of models given the observed data. The key constituent of this posterior probability will be the likelihood of the data given the model, integrating over all parameters in the model. In our case, the data are the words in the corpus, $\mathbf{w}$, and the model is specified by the number of topics, T, so we wish to compute the likelihood $P(\mathbf{w}|T)$. The complication is that this requires summing over all possible assignments of words to topics $\mathbf{z}$. However, we can approximate $P(\mathbf{w}|T)$ by taking the harmonic mean of a set of values of $P(\mathbf{w}|\mathbf{z}, T)$ when $\mathbf{z}$ is sampled from the posterior $P(\mathbf{z}|\mathbf{w}, T)$ (15). Our Gibbs sampling algorithm provides such samples, and the value of $P(\mathbf{w}|\mathbf{z},T)$ can be computed from Eq. **2**.

The Topics of Science

The algorithm outlined above can be used to find the topics that account for the words used in a set of documents. We applied this algorithm to the abstracts of papers published in PNAS from 1991 to 2001, with the aim of discovering some of the topics addressed by scientific research. We first used Bayesian model selection to identify the number of topics needed to best account for the structure of this corpus, and we then conducted a detailed analysis with the selected number of topics. Our detailed analysis involved examining the relationship between the topics discovered by our algorithm and the class designations supplied by PNAS authors, using topic dynamics to identify "hot topics" and using the topic assignments to highlight the semantic content in abstracts.

How Many Topics? To evaluate the consequences of changing the number of topics T, we used the Gibbs sampling algorithm outlined in the preceding section to obtain samples from the posterior distribution over $\mathbf{z}$ at several choices of T. We used all 28,154 abstracts published in PNAS from 1991 to 2001, with each of these abstracts constituting a single document in the corpus (we will use the words abstract and document interchangeably from this point forward). Any delimiting character, including hyphens, was used to separate words, and we deleted any words that occurred in less than five abstracts or belonged to a standard "stop" list used in computational linguistics, including numbers, individual characters, and some function words. This gave us a vocabulary of 20,551 words, which occurred a total of 3,026,970 times in the corpus.

For all runs of the algorithm, we used $\beta = 0.1$ and $\alpha = 50/T$, keeping constant the sum of the Dirichlet hyperparameters, which can be interpreted as the number of virtual samples contributing to the smoothing of θ. This value of β is relatively small and can be expected to result in a fine-grained decomposition of the corpus into topics that address specific research areas. We computed an estimate of $P(\mathbf{w}|T)$ for T values of 50, 100, 200, 300, 400, 500, 600, and 1,000 topics. For all values of T, except the last, we ran eight Markov chains, discarding the first 1,000 iterations, and then took 10 samples from each chain at a lag of 100 iterations. In all cases, the log-likelihood values stabilized within a few hundred iterations, as in Fig. 2. The simulation with 1,000 topics was more time-consuming, and thus we used only six chains, taking two samples from each chain after 700 initial iterations, again at a lag of 100 iterations.

Estimates of $P(\mathbf{w}|T)$ were computed based on the full set of samples for each value of T and are shown in Fig. 3. The results suggest that the data are best accounted for by a model incorporating 300 topics. $P(\mathbf{w}|T)$ initially increases as a function of T, reaches a peak at $T = 300$, and then decreases thereafter. This kind of profile is often seen when varying the dimensionality of a statistical model, with the optimal model being rich enough to fit the information available in the data, yet not so complex as to begin fitting noise. As mentioned above, the value of T found through this procedure depends on the choice of α and β, and it will also be affected by specific decisions made in forming the dataset, such as the use of a stop list and the inclusion of documents from all PNAS classifications. By using just $P(\mathbf{w}|T)$ to choose a value of T, we are assuming very weak prior constraints on the number of topics. $P(\mathbf{w}|T)$ is just the likelihood term in the inference to $P(T|\mathbf{w})$, and the prior $P(T)$ might overwhelm this likelihood if we had a particularly strong preference for a smaller number of topics.

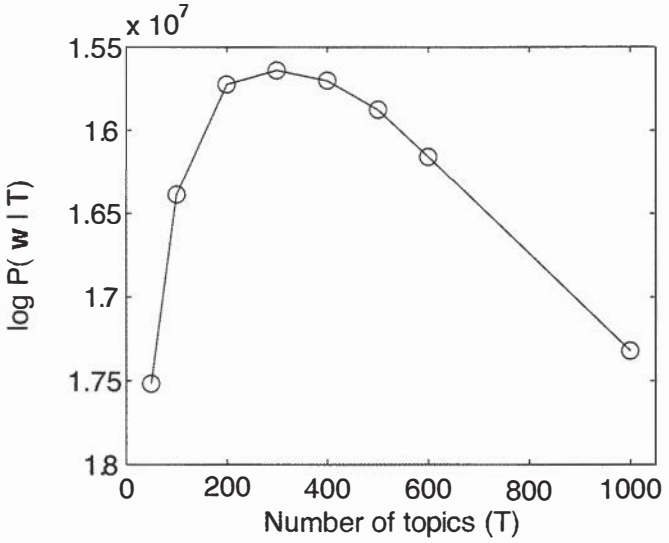


Fig. 3. Model selection results, showing the log-likelihood of the data for different settings of the number of topics, T. The estimated standard errors for each point were smaller than the plot symbols.

Scientific Topics and Classes. When authors submit a paper to PNAS, they choose one of three major categories, indicating whether a paper belongs to the Biological, the Physical, or the Social Sciences, and one of 33 minor categories, such as Ecology, Pharmacology, Mathematics, or Economic Sciences. (Anthropology and Psychology can be chosen as a minor category for papers in both Biological and Social Sciences. We treat these minor categories as distinct for the purposes of our analysis.) Having a class designation for each abstract in the corpus provides two opportunities. First, because the topics recovered by our algorithm are purely a consequence of the statistical structure of the data, we can evaluate whether the class designations pick out differences between abstracts that can be expressed in terms of this statistical structure. Second, we can use the class designations to illustrate how the distribution over topics can reveal relationships between documents and between document classes.

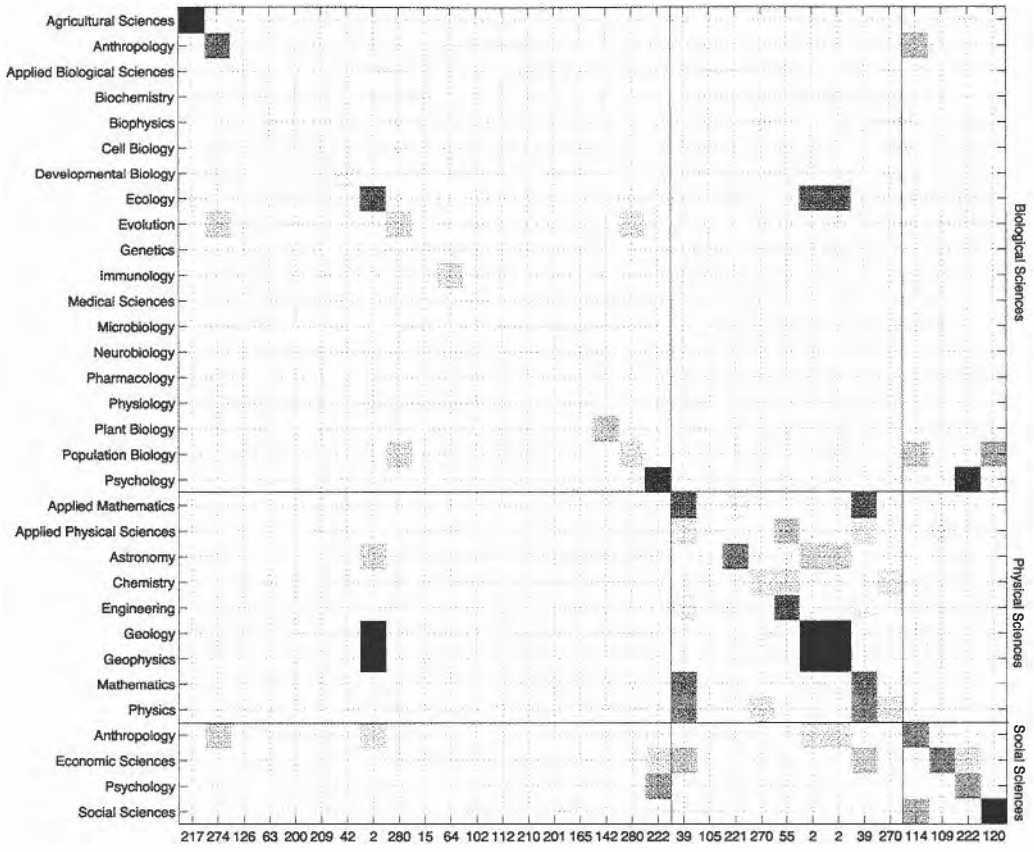
We used a single sample taken after 2,000 iterations of Gibbs sampling and computed estimates of $\theta^{(d)}$ by means of Eq. **7**. (In this and other analyses, similar results were obtained by examining samples across multiple chains, up to the permutation of topics, and the choice of this particular sample to display the results was arbitrary.) Using these estimates, we computed a mean θ vector for each minor category, considering just the 2,620 abstracts published in 2001. We then found the most diagnostic topic for each minor category, defined to be the topic j for which the ratio of θ_j for that category to the sum of θ_j across all other categories was greatest. The results of this analysis are shown in Fig. 4. The matrix shown in Fig. 4 *Upper* indicates the mean value of θ for each minor category, restricted to the set of most diagnostic topics. The strong diagonal is a consequence of our selection procedure, with diagnostic topics having high probability within the classes for which they are diagnostic, but low probability in other classes. The off-diagonal elements illustrate the relationships between classes, with similar classes showing similar distributions across topics.

The distributions over topics for the different classes illustrate how this statistical model can capture similarity in the semantic content of documents. Fig. 4 reveals relationships between specific minor categories, such as Ecology and Evolution, and some of the correspondences within major categories; for example, the minor categories in the Physical and Social Sciences show much greater commonality in the topics appearing in their abstracts than do the Biological Sciences. The results can also be used to assess how much different disciplines depend on particular methods. For example, topic 39, relating to mathematical methods, receives reasonably high probability in Applied Mathematics, Applied Physical Sciences, Chemistry, Engineering, Mathematics, Physics, and Economic Sciences, suggesting that mathematical theory is particularly relevant to these disciplines.

The content of the diagnostic topics themselves is shown in Fig. 4 *Lower*, listing the five words given highest probability by each topic. In some cases, a single topic was the most diagnostic for several classes: topic 2, containing words relating to global climate change, was diagnostic of Ecology, Geology, and Geophysics; topic 280, containing words relating to evolution and natural selection, was diagnostic of both Evolution and Population Biology; topic 222, containing words relating to cognitive neuroscience, was diagnostic of Psychology as both a Biological and a Social Science; topic 39, containing words relating to mathematical theory, was diagnostic of both Applied Mathematics and Mathematics; and topic 270, containing words having to do with spectroscopy, was diagnostic of both Chemistry and Physics. The remaining topics were each diagnostic of a single minor category and, in general, seemed to contain words relevant to enquiry in that discipline. The only exception was topic 109, diagnostic of Economic Sciences, which contains words generally relevant to scientific research. This may be a consequence of the relatively small number of documents in this class (only three in 2001), which makes the estimate of θ extremely unreliable. Topic 109 also serves to illustrate that not all of the topics found by the algorithm correspond to areas of research; some of the topics picked out scientific words that tend to occur together for other reasons, like those that are used to describe data or those that express tentative conclusions.

Finding strong diagnostic topics for almost all of the minor categories suggests that these categories have differences that can be expressed in terms of the statistical structure recovered by our algorithm. The topics discovered by the algorithm are found in a completely unsupervised fashion, using no information except the distribution of the words themselves, implying that the minor categories capture real differences in the content of abstracts, at the level of the words used by authors. It also shows that this algorithm finds genuinely informative structure in the data, producing topics that connect with our intuitive understanding of the semantic content of documents.

Hot and Cold Topics. Historians, sociologists, and philosophers of science and scientists themselves recognize that topics rise and fall in the amount of scientific interest they generate, although whether this is the result of social forces or rational scientific practice is the subject of debate (e.g., refs. 16 and 17). Because our analysis reduces a corpus of scientific documents to a set of topics, it is straightforward to analyze the dynamics of these topics as a means of gaining insight into the dynamics of science. If understanding these dynamics is the goal of our analysis, we can formulate more sophisticated generative models that incorporate parameters describing the change in the prevalence of topics over time. Here, we present a basic analysis based on a post hoc examination of the estimates of θ produced by the model. Being able to identify the "hot topics" in science at a particular point is one of the most attractive applications of this kind of model, providing quantitative measures of the prevalence of particular kinds of research that may be useful for historical purposes and for determination of targets for scientific funding. Analysis at the level of topics provides the opportunity to combine information about the occurrences of a set of semantically related words with cues that come from the content of the remainder of the document, potentially highlighting trends

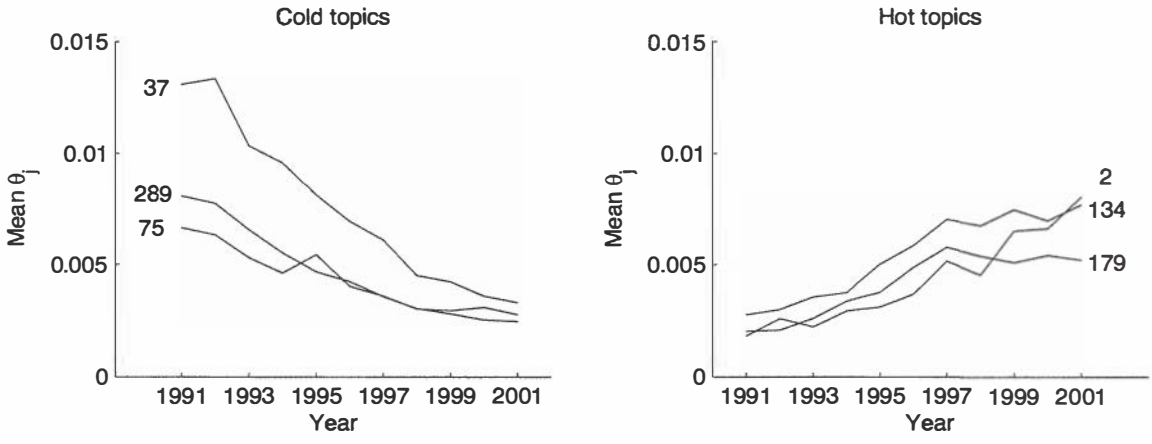


217	274	126	63	200	209
INSECT	SPECIES	GENE	STRUCTURE	FOLDING	NUCLEAR
MYB	PHYLOGENETIC	VECTOR	ANGSTROM	NATIVE	NUCLEUS
PHEROMONE	EVOLUTION	VECTORS	CRYSTAL	PROTEIN	LOCALIZATION
LENS	EVOLUTIONARY	EXPRESSION	RESIDUES	STATE	CYTOPLASM
LARVAE	SEQUENCES	TRANSFER	STRUCTURES	ENERGY	EXPORT
42	**2**	**280**	**15**	**64**	**102**
NEURAL	SPECIES	SPECIES	CHROMOSOME	CELLS	TUMOR
DEVELOPMENT	GLOBAL	SELECTION	REGION	CELL	CANCER
DORSAL	CLIMATE	EVOLUTION	CHROMOSOMES	ANTIGEN	TUMORS
EMBRYOS	CO2	GENETIC	KB	LYMPHOCYTES	HUMAN
VENTRAL	WATER	POPULATIONS	MAP	CD4	CELLS
112	**210**	**201**	**165**	**142**	**222**
HOST	SYNAPTIC	RESISTANCE	CHANNEL	PLANTS	CORTEX
BACTERIAL	NEURONS	RESISTANT	CHANNELS	PLANT	BRAIN
BACTERIA	POSTSYNAPTIC	DRUG	VOLTAGE	ARABIDOPSIS	SUBJECTS
STRAINS	HIPPOCAMPAL	DRUGS	CURRENT	TOBACCO	TASK
SALMONELLA	SYNAPSES	SENSITIVE	CURRENTS	LEAVES	AREAS
39	**105**	**221**	**270**	**55**	**114**
THEORY	HAIR	LARGE	TIME	FORCE	POPULATION
TIME	MECHANICAL	SCALE	SPECTROSCOPY	SURFACE	POPULATIONS
SPACE	MB	DENSITY	NMR	MOLECULES	GENETIC
GIVEN	SENSORY	OBSERVED	SPECTRA	SOLUTION	DIVERSITY
PROBLEM	EAR	OBSERVATIONS	TRANSFER	SURFACES	ISOLATES
		109	**120**		
		RESEARCH	AGE		
		NEW	OLD		
		INFORMATION	AGING		
		UNDERSTANDING	LIFE		
		PAPER	YOUNG		

Fig. 4. (*Upper*) Mean values of θ at each of the diagnostic topics for all 33 PNAS minor categories, computed by using all abstracts published in 2001. Higher probabilities are indicated with darker cells. (*Lower*) The five most probable words in the topics themselves listed in the same order as on the horizontal axis in *Upper*.

that might be less obvious in analyses that consider only the frequencies of single words.

To find topics that consistently rose or fell in popularity from 1991 to 2001, we conducted a linear trend analysis on θ_j by year, using the same single sample as in our previous analyses. We applied this analysis to the sample used to generate Fig. 4. Consistent with the idea that science shows strong trends, with topics rising and falling regularly in popularity, 54 of the topics showed a statistically significant increasing linear trend, and 50 showed a statistically significant decreasing linear trend, both at the $P = 0.0001$ level. The three hottest and coldest topics, assessed by the size of the linear trend test statistic, are shown in Fig. 5. The hottest topics discovered through this analysis were topics 2, 134, and 179, corresponding to global warming and



37	289	75	2	134	179
CDNA	KDA	ANTIBODY	SPECIES	MICE	APOPTOSIS
AMINO	PROTEIN	ANTIBODIES	GLOBAL	DEFICIENT	DEATH
SEQUENCE	PURIFIED	MONOCLONAL	CLIMATE	NORMAL	CELL
ACID	MOLECULAR	ANTIGEN	CO2	GENE	INDUCED
PROTEIN	MASS	IGG	WATER	NULL	BCL
ISOLATED	CHROMATOGRAPHY	MAB	ENVIRONMENTAL	MOUSE	CELLS
ENCODING	POLYPEPTIDE	SPECIFIC	YEARS	TYPE	APOPTOTIC
CLONED	GEL	EPITOPE	MARINE	HOMOZYGOUS	CASPASE
ACIDS	SDS	HUMAN	CARBON	ROLE	FAS
IDENTITY	BAND	MABS	DIVERSITY	KNOCKOUT	SURVIVAL
CLONE	APPARENT	RECOGNIZED	OCEAN	DEVELOPMENT	PROGRAMMED
EXPRESSED	LABELED	SERA	EXTINCTION	GENERATED	MEDIATED

Fig. 5. The plots show the dynamics of the three hottest and three coldest topics from 1991 to 2001, defined as those topics that showed the strongest positive and negative linear trends. The 12 most probable words in those topics are shown below the plots.

climate change, gene knockout techniques, and apoptosis (programmed cell death), the subject of the 2002 Nobel Prize in Physiology. The cold topics were not topics that lacked prevalence in the corpus but those that showed a strong decrease in popularity over time. The coldest topics were 37, 289, and 75, corresponding to sequencing and cloning, structural biology, and immunology. All these topics were very popular in about 1991 and fell in popularity over the period of analysis. The Nobel Prizes again provide a good means of validating these trends, with prizes being awarded for work on sequencing in 1993 and immunology in 1989.

Tagging Abstracts. Each sample produced by our algorithm consists of a set of assignments of words to topics. We can use these assignments to identify the role that words play in documents. In particular, we can tag each word with the topic to which it was assigned and use these assignments to highlight topics that are particularly informative about the content of a document. The abstract shown in Fig. 6 is tagged with topic labels as superscripts. Words without superscripts were not included in the vocabulary supplied to the model. All assignments come from the same single sample as used in our previous analyses, illustrating the kind of words assigned to the evolution topic discussed above (topic 280).

This kind of tagging is mainly useful for illustrating the content of individual topics and how individual words are assigned, and it was used for this purpose in ref. 1. It is also possible to use the results of our algorithm to highlight conceptual content in other ways. For example, if we integrate across a set of samples, we can compute a probability that a particular word is assigned to the most prevalent topic in a document. This probability provides a graded measure of the importance of a word that uses information from the full set of samples, rather than a discrete measure computed from a single sample. This form of highlighting is used to set the contrast of the words shown in Fig. 6 and picks out the words that determine the topical content of the document. Such methods might provide a means of increasing the efficiency of searching large document databases, in particular, because it can be modified to indicate words belonging to the topics of interest to the searcher.

Conclusion

We have presented a statistical inference algorithm for Latent Dirichlet Allocation (1), a generative model for documents in

A generalized[3] fundamental[146] theorem[267] of natural[280] selection[280] is derived[283] for populations[280] incorporating[149] both genetic[280] and cultural[280] transmission[25]. The phenotype[3] is determined[17] by an arbitrary[3] number[287] of multiallelic[3] loci[3] with two[271]-factor[60] epistasis[280] and an arbitrary[149] linkage[3] map[3], as well as by cultural[280] transmission[25] from the parents[280]. Generations[280] are discrete[60] but partially[253] overlapping[146], and mating[280] may be nonrandom[280] at either the genotypic[280] or the phenotypic[280] level[199] (or both). I show[25] that cultural[280] transmission[25] has several[178] important[173] implications[17] for the evolution[280] of population[280] fitness[280], most notably[260] that there is a time[72] lag[72] in the response[213] to selection[280] such that the future[257] evolution[280] depends[103] on the past selection[280] history[280] of the population[280].

Fig. 6. A PNAS abstract (18) tagged according to topic assignment. The superscripts indicate the topics to which individual words were assigned in a single sample, whereas the contrast level reflects the probability of a word being assigned to the most prevalent topic in the abstract, computed across samples.

which each document is viewed as a mixture of topics, and have shown how this algorithm can be used to gain insight into the content of scientific documents. The topics recovered by our algorithm pick out meaningful aspects of the structure of science and reveal some of the relationships between scientific papers in different disciplines. The results of our algorithm have several interesting applications that can make it easier for people to understand the information contained in large knowledge domains, including exploring topic dynamics and indicating the role that words play in the semantic content of documents.

The results we have presented use the simplest model of this kind and the simplest algorithm for generating samples. In future research, we intend to extend this work by exploring both more complex models and more sophisticated algorithms. Whereas in this article we have focused on the analysis of scientific documents, as represented by the articles published in PNAS, the methods and applications we have presented are relevant to a variety of other knowledge domains. Latent Dirichlet Allocation is a statistical model that is appropriate for any collection of documents, from e-mail records and newsgroups to the entire World Wide Web. Discovering the topics underlying the structure of such datasets is the first step to being able to visualize their content and discover meaningful trends.

We thank Josh Tenenbaum, Dave Blei, and Jun Liu for thoughtful comments that improved this paper, Kevin Boyack for providing the PNAS class designations, Shawn Cokus for writing the random number generator, and Tom Minka for writing the code used for the comparison of algorithms. Several simulations were performed on the BlueHorizon supercomputer at the San Diego Supercomputer Center. This work was supported by funds from the NTT Communication Sciences Laboratory (Japan) and by a Stanford Graduate Fellowship (to T.L.G.).

1. Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003) *J. Machine Learn. Res.* **3,** 993–1022.
2. Hofmann, T. (2001) *Machine Learn. J.* **42,** 177–196.
3. Cohn, D. & Hofmann, T. (2001) in *Advances in Neural Information Processing Systems 13* (MIT Press, Cambridge, MA), pp. 430–436.
4. Iyer, R. & Ostendorf, M. (1996) in *Proceedings of the International Conference on Spoken Language Processing* (Applied Science & Engineering Laboratories, Alfred I. duPont Inst., Wilmington, DE), Vol 1., pp. 236–239.
5. Bigi, B., De Mori, R., El-Beze, M. & Spriet, T. (1997) in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings* (IEEE, Piscataway, NJ), pp. 535–542.
6. Ueda, N. & Saito, K. (2003) in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA), Vol. 15.
7. Erosheva, E. A. (2003) in *Bayesian Statistics* (Oxford Univ. Press, Oxford), Vol. 7.
8. Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) *J. R. Stat. Soc. B* **39,** 1–38.
9. Minka, T. & Lafferty, J. (2002) Expectation-propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence* (Elsevier, New York).
10. Newman, M. E. J. & Barkema, G. T. (1999) *Monte Carlo Methods in Statistical Physics* (Oxford Univ. Press, Oxford).
11. Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice* (Chapman & Hall, New York).
12. Liu, J. S. (2001) *Monte Carlo Strategies in Scientific Computing* (Springer, New York).
13. Geman, S. & Geman, D. (1984) *IEEE Trans. Pattern Anal. Machine Intelligence* **6,** 721–741.
14. Manning, C. D. & Schutze, H. (1999) *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, MA).
15. Kass, R. E. & Raftery, A. E. (1995) *J. Am. Stat. Assoc.* **90,** 773–795.
16. Kuhn, T. S. (1970) *The Structure of Scientific Revolutions* (Univ. of Chicago Press, Chicago), 2nd Ed.
17. Salmon, W. (1990) in *Scientific Theories, Minnesota Studies in the Philosophy of Science,* ed. Savage, C. W. (Univ. of Minnesota Press, Minneapolis), Vol. 14.
18. Findlay, C. S. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 4874–4876.

Mapping subsets of scholarly information

Paul Ginsparg[†‡], Paul Houle, Thorsten Joachims, and Jae-Hoon Sul

Cornell University, Ithaca, NY 14853

We illustrate the use of machine learning techniques to analyze, structure, maintain, and evolve a large online corpus of academic literature. An emerging field of research can be identified as part of an existing corpus, permitting the implementation of a more coherent community structure for its practitioners.

The arXiv (see http://arXiv.org; for general background, see ref. 1) is an automated repository of >250,000 full-text research articles (as of mid-October 2003) in physics and related disciplines, going back more than a decade and growing at a rate of 40,000 new submissions per year. It serves >10 million requests per month (2), including tens of thousands of search queries per day and >20 million full-text downloads during 2002. It is a significant example of a Web-based service that has changed the practice of research in a major scientific discipline. It provides nearly comprehensive coverage of large areas of physics and serves as an on-line seminar system for those areas. It also provides a significant resource for model building and algorithmic experiments in mapping scholarly domains. Together with the SLAC SPIRES-HEP database, it provides a public resource of full-text articles and associated citation trees of many millions of links, with a focused disciplinary coverage and rich usage data. [The Stanford Linear Accelerator Center SPIRES-HEP database has comprehensively catalogued the High Energy Particle Physics (HEP) literature online since 1974, and indexes >500,000 high-energy physics-related articles including their full citation tree (see ref. 3).]

In what follows, we use arXiv data to illustrate how machine learning methods can be used to analyze, structure, maintain, and evolve a large online corpus of academic literature. The specific application will be to train a support vector machine text classifier to extract an emerging research area from a larger-scale resource. The automated detection of such subunits can play an important role in disentangling other subnetworks and associated subcommunities from the global network. Our notion of "mapping" here is in the mathematical sense of associating attributes to objects, rather than in the sense of visualization tools. Although the former underlies the latter, we expect the two will increasingly go hand-in-hand (for a recent review, see ref. 4).

Text Classification

The goal of text classification is the automatic assignment of documents to a fixed number of semantic categories. In the "multilabel" setting, each document can be in zero or one or more categories. Efficient automated techniques are essential to avoid tedious and expensive manual category assignment for large document sets. A "knowledge-engineering" approach, involving hand-crafting accurate text-classification rules, is surprisingly difficult and time-consuming (5). We therefore take a machine learning approach to generating text-classification rules automatically from examples.

The machine learning approach can be phrased as a supervised learning problem. The learning task is represented by the training sample S_n

$$(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n) \qquad [1]$$

of size n documents, where $\vec{x}_i$ represents the document content. In the multilabel setting, each category label is treated as a separate binary classification problem. For each such binary task, $y_i \in \{-1,+1\}$ indicates whether a document belongs to a particular class. The task of the learning algorithm, L, is to find a decision rule, $h_L:\vec{x} \to \{-1,+1\}$, based on S_n that classifies new documents, x, as accurately as possible.

Documents need to be transformed into a representation suitable for the learning algorithm and the classification task. Information retrieval research suggests that words work well as representation units and that for many tasks their ordering can be ignored without losing too much information. This type of representation is commonly called the "bag-of-words" model, an attribute-value representation of text. Each text document is represented by a vector in the lexicon space, i.e., by a "term frequency" (TF) feature vector $\mathrm{TF}(w_i, x)$, with component values equal to the number of times each distinct word, w_i, in the corpus occurs in the document x. Fig. 1 shows an example feature vector for a particular document.

This basic representation is ordinarily refined in a few ways.

TF × Inverse Document Frequency (IDF) Weighting. Scaling the components of the feature vector with their $\mathrm{IDF}(w_i)$ (6) often leads to improved performance. In general, $\mathrm{IDF}(w_i)$ is some decreasing function of the word frequency $\mathrm{DF}(w_i)$, equal to the number of documents in the corpus which contain the word, w_i. For example,

$$\mathrm{IDF}(w_i) = \log\left(\frac{n}{\mathrm{DF}(w_i)}\right), \qquad [2]$$

where n is the total number of documents. Intuitively, the IDF assumes that rarer terms have more significance for classification purposes and hence gives them greater weight. To compensate for the effect of different document lengths, each document feature vector $\vec{x}_i$ is normalized to unit length: $\|\vec{x}_i\| = 1$.

Stemming. Instead of treating each occurrence form of a word as a different feature, stemming is used to project the different forms of a word onto a single feature, the word stem, by removing inflection information (7). For example "computes," "computing," and "computer" are all mapped to the same stem "comput." The terms "word" and "word stem" are used synonymously in the following text.

Stopword Removal. For many classification tasks, common words like "the," "and," or "he" do not help discriminate between

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

Abbreviations: SVM, support vector machine; TF, term frequency; IDF, inverse document frequency.

[†]Presenter of invited talk at Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

[‡]To whom correspondence should be addressed. E-mail: ginsparg@cornell.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0308253100

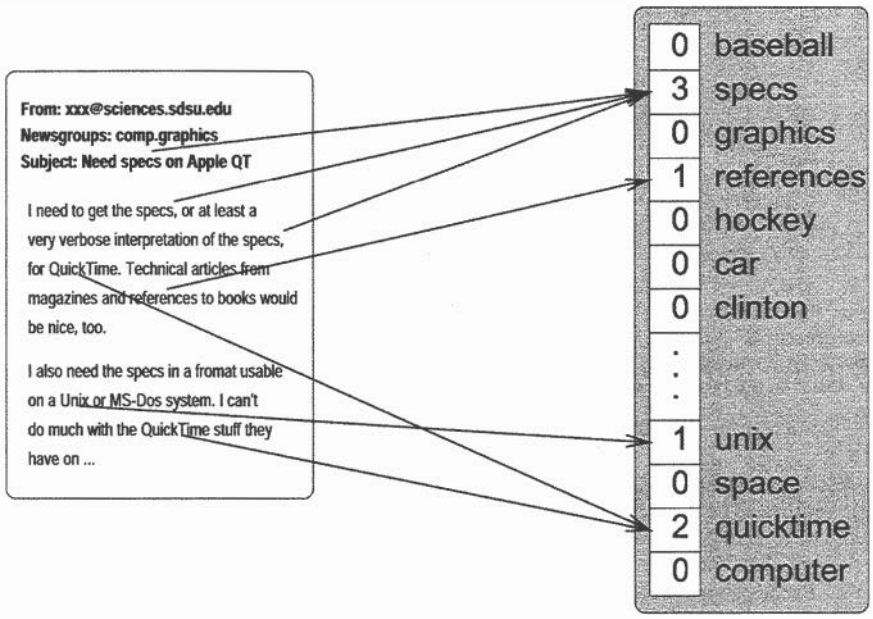


Fig. 1. Representing text as a feature vector.

document classes. Stopword removal describes the process of eliminating such words from the document by matching against a predefined list of stopwords. We use a standard stoplist of ≈300 words.

Support Vector Machines (SVMs)

SVMs were developed by Vapnik and coworkers (8) based on the structural risk minimization principle from statistical learning theory. They have proven to be a highly effective method for learning text-classification rules, achieving state-of-the-art performance on a broad range of tasks (9, 10). Two main advantages of using SVMs for text classification lie in their ability to handle the high-dimensional feature spaces arising from the bag-of-words representation. From a statistical perspective, they are robust to overfitting and are well suited for the statistical properties of text. From a computational perspective, they can be trained efficiently despite the large number of features. A detailed overview of the SVM approach to text classification, with more details on the notation used below, is given in ref. 11.

In their basic form, SVMs learn linear decision rules,

$$h(\vec{x}) = \text{sgn}\{\vec{w}\cdot\vec{x} + b\}, \qquad [3]$$

described by a weight vector $\vec{w}$ and a threshold b, from an input sample of n training examples, $S_n = ((\vec{x}_1, y_1), \cdots, (\vec{x}_n, y_n))$, $\vec{x}_1 \in \mathbb{R}^N, y_i \in \{-1, +1\}$. For a linearly separable S_n, the SVM finds the hyperplane with maximum Euclidean distance to the closest training examples. This distance is called the margin δ, as depicted in Fig. 2. Geometrically, the hyperplane is defined by its normal vector, $\vec{w}$, and its distance from the origin, $-b$. For nonseparable training sets, the amount of training error is measured by using slack variables, ξ_i.

Computing the position of the hyperplane is equivalent to solving the following convex quadratic optimization problem (8):

Optimization Problem 1 [SVM (Primal)].

$$\text{minimize:} \quad V(\vec{w}, b, \vec{\xi}) = \frac{1}{2}\vec{w}\cdot\vec{w} + C\sum_{i=1}^{n}\xi_i \qquad [4]$$

$$\text{subject to:} \quad \forall_{i=1}^{n}: y_i[\vec{w}\cdot\vec{x}_i + b] \geq 1 - \xi_i \qquad [5]$$

$$\forall_{i=1}^{n}: \xi_i > 0. \qquad [6]$$

The margin of the resulting hyperplane is $\delta = 1/\|\vec{w}\|$.

The constraints (Eq. **5**) require that all training examples are classified correctly up to some slack ξ_i. If a training example lies on the "wrong" side of the hyperplane, we have the corresponding $\xi_i \geq 1$, and thus $\sum_{i=1}^{n}\xi_i$ is an upper bound on the number of training errors. The factor C in Eq. **4** is a parameter that allows trading off training error vs. model complexity. The optimal value of this parameter depends on the particular classification task and must be chosen by means of cross-validation or by some other model selection strategy. For text classification, however, the default value of $C = 1/\max_i\|\vec{x}_i\|^2 = 1$ has proven to be effective across a large range of tasks (11).

OP1 has an equivalent dual formulation:

Optimization Problem 2 [SVM (Dual)].

$$\text{maximize:} \quad W(\vec{\alpha}) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}y_i y_j \alpha_i \alpha_j (\vec{x}_i\cdot\vec{x}_j) \qquad [7]$$

$$\text{subject to:} \quad \sum_{i=1}^{n}y_i\alpha_i = 0 \qquad [8]$$

$$\forall_i \in [1..n]: 0 \leq \alpha_i \leq C. \qquad [9]$$

From the solution of the dual, the classification rule solution can be constructed as

$$\vec{w} = \sum_{i=1}^{n}\alpha_i y_i \vec{x}_i \quad \text{and} \quad b = y_{usv} - \vec{w}\cdot\vec{x}_{usv}, \qquad [10]$$

where $(\vec{x}_{usv}, y_{usv})$ is some training example with $0 < \alpha_{usv} < C$. For the experiments in this article, SVM$^{\text{light}}$ (11) is used for solving the dual-optimization problem. (SVM$^{\text{light}}$ is available at http://svmlight.joachims.org/.) More detailed introductions to SVMs can be found in refs. 12 and 13.

An alternative to the approach here would be to use Latent Semantic Analysis (14) to generate the feature vectors. Latent Semantic Analysis can potentially give better recall by capturing

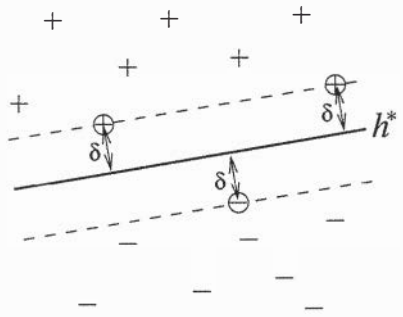


Fig. 2. A binary classification problem (+ vs. −) in two dimensions. The hyperplane h^* separates positive and negative training examples with maximum margin δ. The examples closest to the hyperplane are called "support vectors" (marked with circles).

partial synonymy in the word representation, i.e., by bundling related words into a single feature. The reduction in dimension can also be computationally efficient, facilitating capture of the full-text content of papers, including their citation information. On the other hand, the SVM already scales well to a large feature space and performs well at determining relevant content through suitable choices of weights. On a sufficiently large training set, the SVM might even benefit from access to fine distinctions between words potentially obscured by the latent semantic analysis bundling. It is thus an open question, well worth further investigation, as to whether latent semantic analysis applied to full text could improve performance of the SVM in this application without loss of computational efficiency. Generative models for text classification provide yet another alternative approach, in which each document is viewed as a mixture of topics, as in the statistical inference algorithm for Latent Dirichlet Allocation used in ref. 15. This approach can, in principle, provide significant additional information about document content but does not scale well to a large document corpus, so the real-time applications intended here would not yet be computationally feasible in that framework.

arXiv Benchmarks

Before using the machine learning framework to identify new subject area content, we first assessed its performance on the existing (author-provided) category classifications. Roughly 180,000 titles and abstracts were fed to model-building software, which constructed a lexicon of ≈100,000 distinct words and produced training files containing the TD×IDF document vectors for SVM^{light}. [Although the SVM machinery could easily be used to analyze the full document content, previous experiments (11) suggest that well written titles and abstracts provide a highly focused characterization of content at least as effective for our document classification purposes.] The set of support vectors and weight parameters output by SVM^{light} was converted into a form specific to the linear SVM, Eq. **3**: a weight vector $\vec{w}_c$ and a threshold b_c, where c is an index over the categories.

As seen in Fig. 3, the success of the SVM in classifying documents improves as the size of a category increases. The SVM is remarkably successful at identifying documents in large (>10,000 documents) categories and less successful at identifying smaller subject areas (<500 documents). A cutoff was imposed to exclude subject areas with <100 documents. (Some of the smaller subject areas are known to be less topically focused, so the difficulty in recall, based solely on title/abstract terminology, was expected.)

In experiments with small categories (N < 1,000), the SVM consistently chose thresholds that resulted in unacceptably low recall (i.e., missed documents relevant to the category). This is in part because the null hypothesis, that no documents are in the

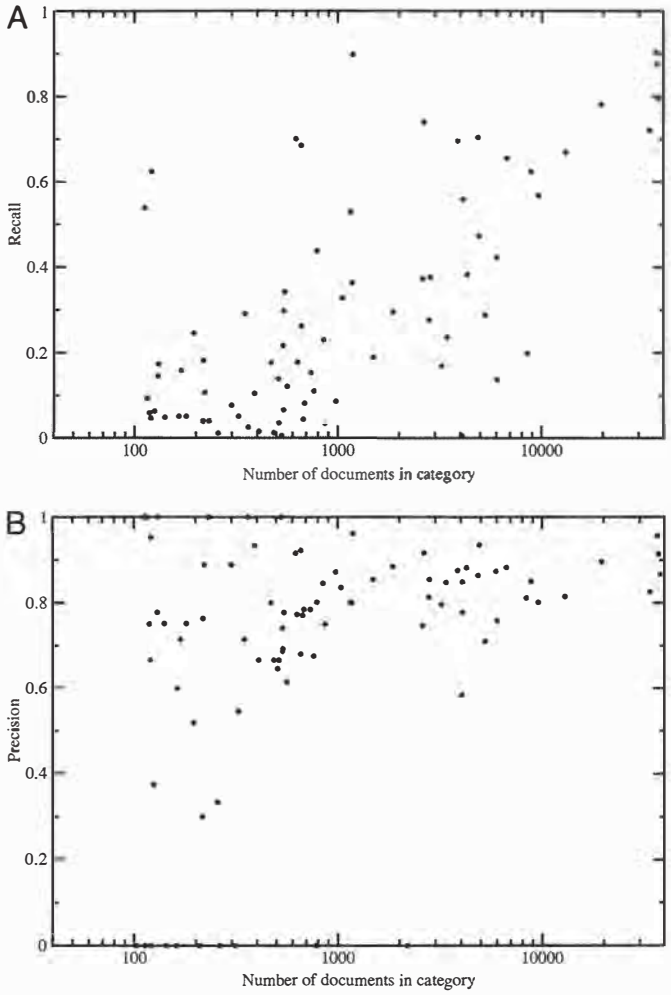


Fig. 3. Recall and precision as functions of category size for 78 arXiv major categories and minor subject classes. Two thirds of the sample was used as a training set and one third as a test set. The four largest categories, each with >30,000 documents are cond-mat, astro-ph, hep-th, and hep-ph.

category, describes the data well for a small category; e.g., if only 1,000 of 100,000 documents are in a category, the null hypothesis has a 99% accuracy. ("Accuracy" is the percentage of documents correctly classified. We also employ other common terminology from information retrieval: "precision" is the fraction of those documents retrieved that are relevant to a query, and "recall" is the fraction of all relevant documents in the collection that are retrieved.) To counteract the bias against small categories, 10% of the training data was used as a validation set to fit a sigmoid probability distribution $1/[1 + \exp(Af + B)]$ (16). This converts the dot product $\vec{x}_i \cdot \vec{w}_c$ between document vector and weight vector into a probability, $P(i \in c \mid x_i)$, that document i is a member of category c, given its feature vector x_i. Use of the probability in place of the uncalibrated signed distance from the hyperplane output by the SVM permitted greater levels of recall for small categories.

Other experiments showed that the use of TF×IDF weighting as in Eq. **2** improved accuracy consistently over pure TF weighting, so TF×IDF weighting was used in the experiments to follow. We also used a document frequency threshold to exclude rare words from the lexicon but found little difference in accuracy between using a document occurrence threshold of two and five. (Words that appeared in fewer than two documents constituted ≈50% of the lexicon, and those that appeared in

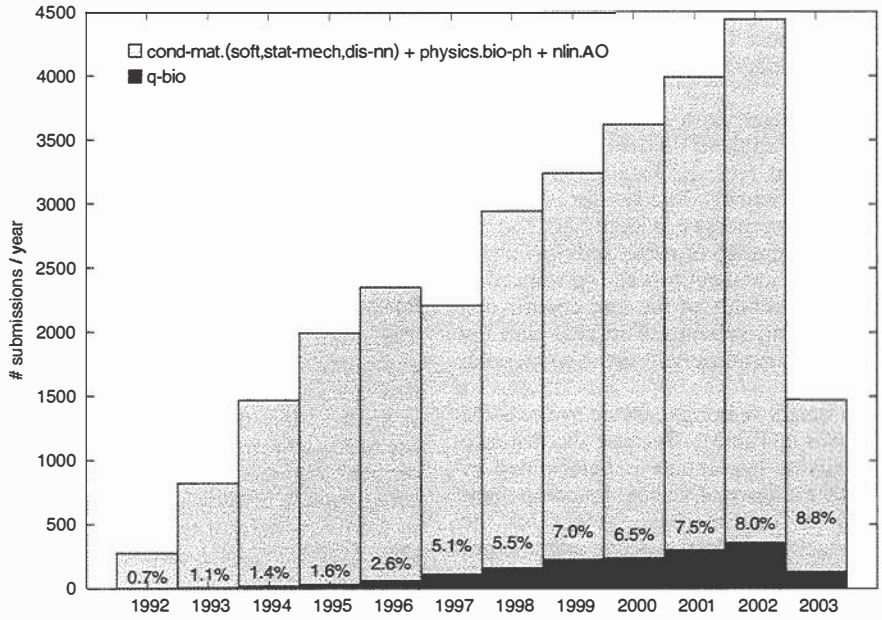


Fig. 4. The number of submissions per year from 1992 through April 2003 in the particular subsets of arXiv subject areas of cond-mat, physics, and nlin most likely to have "quantitative biology" content. The percentage of q-bio content in these areas grew from ≈1% to nearly 10% during this timeframe, suggesting a change in intellectual activity among arXiv-using members of these communities.

fewer than five documents, ≈70%. Ignoring rare and consequently uninformative words hence reduces the computational needs.) Increasing the weight of title words with respect to abstract words, on the other hand, consistently worsened accuracy, indicating that words in a well written abstract contain as much or more content classification import as words in the title. Changes in the word tokenization and stemming algorithms did not have a significant impact on overall accuracy. The default value of $C = 1$ in Eq. **9** was preferred.

q-bio Extraction

Recent anecdotal evidence indicates an intellectual trend among physicists toward work in biology, ranging from biomolecules, molecular pathways and networks, gene expression, cellular and multicellular systems to population dynamics and evolution. This work has appeared in separate parts of the archive, in particular, under "Soft Condensed Matter," "Statistical Mechanics," "Disordered Systems and Neural Networks," "Biophysics," and "Adaptive and Self-Organizing Systems" (abbreviated cond-mat.soft, cond-mat.stat-mech, cond-mat.dis-nn, physics.bio-ph, and nlin.AO). A more coherent forum for the exchange of these ideas was requested, under the nomenclature "Quantitative Biology" (abbreviated "q-bio").

To identify first whether a real trend to nurture and amplify indeed existed and to create a training set, volunteers were enlisted to identify the q-bio content from the subject areas above in which it was most highly focused. Of 5,565 such articles received from January 2002 through March 2003, 466 (8.4%) were found to have one of the biological topics above as its primary focus. The total number of distinct words in these titles, abstracts, plus author names, was 23,558, of which 7,984 were above the DF = 2 document frequency threshold. (Author names were included in the analysis because they have potential "semantic" content in this context, i.e., are potentially useful for document classification. The SVM algorithm will automatically determine whether to use the information by choosing suitable weights.)

A data-cleaning procedure was used, in which SVM^{light} was first run with $C = 10$. We recall from Eqs. **4** and **9** that larger C penalizes training errors and requires larger α values to fit the data. Inspecting the "outlier" documents (17) with the largest $|\alpha_i|$ then permitted manual cleaning of the training set. Ten documents were moved into q-bio, and 15 documents were moved out, for a net movement to 461 q-bio (8.3%) of the 5,565 total. Some of the others flagged involved word confusions, e.g., "genetic algorithms" typically involved programming rather than biological applications. Other q-bio words with frequent non-biological senses were "epidemic" (used for rumor propagation), "evolution" (used also for dynamics of sandpiles), "survival probabilities," and extinction. "Scale-free networks" were sometimes used for biological applications and sometimes not. To

Table 1. The most positive, a few intermediate, and the most negative components of the q-bio classifying weight vector

protein	+8.57	⋮		point	−1.04
dna	+7.08	forward	+0.00	equation	−1.05
biological	+5.06	minimalist	+0.00	boundary	−1.06
neuron	+5.05	region	+0.00	social	−1.06
rna	+3.30	confinement	+0.00	n	−1.09
gene	+3.21	implies	+0.00	relaxation	−1.14
mutation	+3.15	96	+0.00	fluid	−1.15
population	+3.11	y_togashi	+0.00	indian	−1.15
epidemic	+3.05	n_wingreen	+0.00	spin	−1.17
biology	+3.02	mean_free	+0.00	spin_glass	−1.17
disease	+2.93	narrower	+0.00	traffic	−1.18
cell	+2.90	shot	−0.00	system	−1.30
neural	+2.89	repton	−0.00	polymer	−1.33
brain	+2.83	kyoto	−0.00	class	−1.35
ecosystem	+2.56	regular	−0.00	emerge	−1.36
tissue	+2.52	generalization	−0.00	gradient	−1.39
sequence	+2.51	d_saakian	−0.00	quantum	−1.43
genetic	+2.51	conformity	−0.00	surface	−1.43
bacterial	+2.48	aware	−0.00	synchronization	−1.45
blood	+2.43	even_though	−0.00	market	−1.47
genome	+2.37	practitioner	−0.00	particle	−1.52
peptide	+2.37	permittivity	−0.00	polyelectrolyte	−1.53
infection	+2.34	⋮		world	−1.57

help resolve some of these ambiguities, the vocabulary was enlarged to include a list of most frequently used two-word phrases with semantic content different from their constituent words.

With a training set fully representative of the categories in question, it was then possible to run the classifier on the entirety of the same subject area content received from 1992 through April 2003, a total of 28,830 documents. The results are shown in Fig. 4. A total of 1,649 q-bio documents was identified, and the trend toward an increasing percentage of q-bio activity among these arXiv users is evident; individual authors can be tracked as they migrate into the domain. Visibility of the new domain can be further enhanced by referring submitters in real time by means of the automated classifier running behind the submission interface.

Some components of the weight vector generated by the SVM for this training set are shown in Table 1. Because the distance of a document to the classifying hyperplane is determined by taking the dot product with the normal vector, its component values can be interpreted as the classifying weight for the associated words. The approach here illustrates one of the major lessons of the past decade: the surprising power of simple algorithms operating on large datasets.

Although implemented as a passive-dissemination system, the arXiv has also played a social engineering role, with active research users developing an affinity to the system and adjusting their behavior accordingly. They scan new submissions on a daily basis, assume others in their field do so, are consequently aware of anything relevant that has appeared there (whereas anything that doesn't may as well not exist), and use it to stake intellectual priority claims in advance of journal publication. We see further that machine learning tools can characterize a subdomain and thereby help accelerate its growth by the interaction of an information resource with the social system of its practitioners. [Some other implications of these methods, including potential modification of the peer review system, are considered elsewhere (2).]

Postscript

The q-bio extraction described above was not just a thought experiment, but a prelude to an engineering experiment. The new category went online in mid-September 2003 (see http://arXiv.org/new/q-bio.html), at which time past submitters flagged by the system were asked to confirm the q-bio content of their submissions and to send future submissions to the new category. The activity levels at the outset corresponded precisely to the predictions of the SVM text classifier and later began to show indications of growth catalyzed by the public existence of the new category.

This work was supported by National Science Foundation Agreement 0132355 (to P.G.), funding from the Cornell University Library (to P.H.), National Science Foundation Career Award 0237381 (to T.J.), and the National Science Foundation KD-D Project (to T.J.).

1. Ginsparg, P. (2001) in *Electronic Publishing in Science II, Proceedings of Joint ICSU Press/UNESCO Conference, Paris, France* (ICSU Press, Paris) Available at http://users.ox.ac.uk/icsuinfo/ginspargfin.htm and at http://arXiv.org/blurb/pg01unesco.html. Accessed January 22, 2004.
2. Ginsparg, P. (2003) *Sci. Technol. Libraries* **22,** 5–18, preprint available at http://arXiv.org/blurb/pg02pr.html.
3. O'Connell, H. B. (2002) http://arXiv.org/physics/0007040.
4. Borner, K., Chen, C. & Boyack, K. (2003) *Annu. Rev. Inf. Sci. Technol.* **37,** 179–255.
5. Hayes, P. & Weinstein, S. (1990) in *Proceedings of IAAI-90, 2nd Conference on Innovative Applications of Artificial Intelligence*, eds. Rappaport, A. & Smith, R. (AAAI Press, Menlo Park, CA), pp. 49–66.
6. Salton, G. & Buckley, C. (1988) *Inf. Processing Manage.* **24,** 513–523.
7. Porter, M. (1980) *Program (Autom. Libr. Inf. Syst.)* **14,** 130–137.
8. Vapnik, V. (1998) *Statistical Learning Theory* (Wiley, Chichester, U.K.).
9. Dumais, S., Platt, J., Heckerman, D. & Sahami, M. (1998) in *Proceedings of CIKM-98, Seventh ACM International ACM Conference in Information and Knowledge Managment* (Association for Computing Machinery, New York), pp. 148–155.
10. Joachims, T. (1998) in *Proceedings of the European Conference on Machine Learning* (Springer. Berlin), pp. 137–142.
11. Joachims, T. (2002) *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms* (Kluwer, Dordrecht, The Netherlands).
12. Burges, C. (1998) *Data Mining Knowledge Discovery* **2,** 121–167.
13. Cristianini, N. & Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* (Cambridge Univ. Press, NY).
14. Landauer, T. K. & Dumais, S. T. (1997) *Psychol. Rev.* **104,** 211–240.
15. Griffiths, T. L. & Steyvers, M. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5228–5235.
16. Platt, J. (1999) in *Advances in Large Margin Classifiers* (MIT Press, Cambridge, MA), pp. 61–74.
17. Guyon, I., Matic, N. & Vapnik, V. (1996) in *Advances in Knowledge Discovery and Data Mining* (MIT Press, Cambridge, MA; AAAI Press, Menlo Park, CA), pp. 181–203.

A method for finding communities of related genes

Dennis M. Wilkinson and Bernardo A. Huberman[a]

Stanford University and HP Laboratories, 1501 Page Mill Road, Palo Alto, CA 94394

We present a method for creating a network of gene co-occurrences from the literature and partitioning it into communities of related genes. The way in which our method identifies communities makes it likely that the component genes of each community will be related by their function. The method processes a large database of article abstracts, synthesizing information from many sources to shed light on groups of genes that have been shown to interact. It is a tool to be used by researchers in the biomedical sciences to swiftly search for known interactions and to provide insight into unexplored connections. The partitioning procedure is designed to be particularly applicable to large networks in which individual nodes may play a role in more than one community. In this paper, we explain the details of the method, in particular the partitioning process. We also apply the method to produce communities of genes related to colon cancer and show that the results are useful.

The automated analysis of biomedical text is useful in any form, because knowledge in the biomedical sciences is predominantly disseminated in the form of journal articles. However, when applied to the subject of human gene function, automated text analysis is critically important. There are ≈15,000 currently known human genes and >1 million related articles in the Medline database[b] alone. Moreover, genes act in a complex interrelated way, so information from many experiments is necessary to explain the function of a typical gene. A comprehensive study of even a simple cellular process involving several genes might require a researcher to be familiar with hundreds of articles. Merely locating all relevant articles in a database by using a simple search utility would be time consuming, not to mention inefficient and difficult, because of shortcomings of the human gene nomenclature system. In contrast, our method indexes gene symbol occurrences in all articles of large database such as Medline in <1 day[c] and then can produce a list of communities of functionally related genes in another half day.[d]

In this article, we present a method to find communities of related genes. The method creates a network of gene symbol co-occurrences from Medline article abstracts and partitions this network into communities. The genes in each community are likely to be functionally related because of the way in which the communities are identified, and because most recent research on genes and proteins has been devoted to their function. This method can thus be a valuable tool that both summarizes available information and indicates possible directions of research. The format of the results is designed to make them easy to use. The results can easily include a list of the Medline PubMed identification numbers (PMID) for articles containing each gene and pair of genes to facilitate research. Varying the user-selected key words (see *Method Overview*) allows the method to be applied repeatedly and focused on particular topics of interest.

We apply our method to the Medline database to identify communities of genes related to colon cancer. We show that genes placed together in a community that are not explicitly connected in any Medline article or in the Online Mendelian Inheritance in Man (OMIM)[e] listing for either gene can nevertheless be related by their function. The communities thereby imply connections among genes that may otherwise be overlooked or that would require much time and effort to be found manually. We also show that our method separates genes that co-occur but are not functionally related into different communities. Finally, we demonstrate cases in which a node common to two communities indicates a link between two groups of related genes.

It is important to note that the gene communities in the results are not meant to perfectly reproduce biological reality. The communities are simply interesting artifacts within the network that provide a powerful method for organizing and presenting information from the literature.

Method Overview

Gene symbol mentions are first extracted from almost all[f] 12.5 million Medline article titles and abstracts. We then select sets of genes found to be statistically correlated to a set of user-selected (related) key words. These two steps are performed following the procedure of ref. 1. This procedure includes steps to account for alias symbols and to distinguish gene symbol abbreviations from identical abbreviations referring to other concepts[g] (2). Selecting genes correlated to certain key words ensures continuity of biological function of the genes considered and reduces the number of genes considered so the results can be readable and useful.

Networks are then created from these sets of genes. In the networks, each node represents a gene, and an edge connects two genes if they co-occur in at least one article. The degree distribution of the networks follows a power law, as we show, so their clustering structure is scale-free and there is no typical community size. Therefore, to find communities, we partition the graph using a nonlocal process exploiting the concept of betweenness centrality (3).

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

Abbreviations: *COX-2*, cyclooxygenase 2; *PTGS2*, prostaglandin-endoperoxide synthase 2; GN, Girvan–Newman; PMID, PubMed identification number.

[a]To whom correspondence should be addressed. E-mail: huberman@hpl.hp.com.

[b]Medline is the foremost English-language database of biomedical articles. The search utility for Medline is PubMed (www.ncbi.nlm.nih.gov/entrez/query.fcgi).

[c]The machine we used is a standard 1-GHz machine with an Intel Pentium 3 processor running RED HAT LINUX.

[d]The time to perform this step increases as nm^2, where n is the number of genes in the network and m is the number of pairs of genes that co-occur. As we explain later, genes are selected to create a network, and if the network is too large, this step could be very slow. We found that a size of <1,000 genes is generally tractable for our method.

[e]http://www3.ncbi.nlm.nih.gov/omim. This web site provides detailed information about many genes, proteins, and other biological objects as well as references to related articles.

[f]We omitted the small fraction of abstracts published before 1990, because they very rarely discuss gene function and tend to use outmoded nomenclature. In addition, we neglect abstracts that mention more than four genes, because they are typically abstracts of survey-type articles that impair the community identification process. These two types of articles form a very small fraction of the Medline database.

[g]An example is DCC, which may be used to refer to the gene "deleted in colon cancer" or the cancer assay method "dextran-coated charcoal." Such ambiguous symbols are very common because of the frequent use of abbreviations in biological texts.

The partitioning process may be applied to any network, but it is particularly applicable to networks of several hundred to 1,000 nodes in which nodes may play a role in more than one community. It is based on the process of Girvan and Newman (GN) (ref. 4; for a faster algorithm for finding communities, see ref. 5), which was shown to give very good results for a variety of small graphs. The general idea of our process is the same as that of GN, but the details are significantly different. Our modifications allow nodes to be placed in several communities if the structure of the network indicates that the nodes belong there, and they provide a quantitative estimate of how strongly each node belongs to each community. This is important when single nodes play a role in several communities or when the source information is incomplete or flawed. It can also indicate a link between two communities that have one or more nodes in common, and it "smooths" the process of partitioning, which for any large network is somewhat arbitrary. The modifications also allow communities to be identified as discrete units. Identifying discrete communities is particularly useful when community sizes are not known in advance and makes the results easier to use if the network is large (6).

Motivation and Previous Work

For the most part, biologists now understand the rules by which the system of genes, proteins, RNAs, and other cellular constituents operates; what remains is to determine the exact details of this system. A worldwide effort is underway in the biomedical community to identify and understand the cellular interactions at the root of human health. Given the enormous number of human genes and the complex interrelated nature of gene and protein interaction, this task is more than a little daunting, and accomplishing it will involve an unprecedented level of collaboration and information exchange. However, the current condition of knowledge organization in the field makes extensive collaboration and complete information exchange difficult.

As mentioned above, information pertinent to human gene function exists largely in the form of an astoundingly large number of journal articles. Medline yields 1.5 million hits when queried for "gene" or "protein" with "human," ≈150,000 of which were published in 2002 alone. Our results, taking into account co-occurrences within the set of 682 genes we identified as correlated to colon cancer, were created from the 7,985 article abstracts from an astonishing 904[h] different journals. Given these numbers, it is easy to see that an expert, although familiar with many hundreds of articles, could nonetheless be unaware of developments related to his or her area of interest. And, whereas online biomedical databases provide easy access to abstracts, a manual literature survey would encounter difficulties beyond the large number of results, due to the nomenclature system for human genes. Both the existence of multiple alias symbols for many genes and the frequent occurrence of unrelated abbreviations equivalent to gene symbols interfere with any simple search utility.

Despite the impracticality of an exhaustive manual search, online databases of journal abstracts present a gold mine of available information. In fact, the ability to sift through millions of abstracts, extract pertinent information, and present it in a useful format is arguably essential to the understanding of human gene function. Accordingly, automated text analysis has been an area of focus in the field of bioinformatics.

One approach has been to extract detailed information by using natural language-processing techniques (7–17). Our method follows a different line of attack: only simple information, such as gene and protein names, is extracted from each article, and more detailed conclusions are then inferred from this information. Gene and protein term identification in particular has been simplified by the recent appearance of online libraries of gene and protein symbols (refs. 18–20 show this can otherwise be a major task). However, data obtained by simple term matching will be highly error-prone due to false positive identifications of human gene symbols, unless carefully treated.

A reasonable conclusion that can be drawn from gene occurrence data is that genes mentioned in the same article are related in some way. This has been shown to be true both on large (21) and small (22, 23) scales. It is also possible to connect genes to key words found in articles and thus to biological processes, as in refs. 1 and 24. These results have been applied in conjunction with natural language-processing techniques to find related groups of genes, from among a restricted set of genes mentioned in a restricted set of articles, in refs. 25 and 26. Our method, while similar, has a very different way of finding communities that requires neither the preprocessing step of selecting genes or articles nor natural language processing.

Obtaining Co-occurrence Data

As stated above, the first step of the method is to identify literature co-occurrences of genes relevant to a disease by using the procedure of ref. 1. This section is simply a brief summary of this procedure; for more detail, please see the referenced article.

Using a list of all official and alias symbols for human genes compiled from the Human Genome Organisation (HUGO) (www.gene.ucl.ac.uk/nomenclature), OMIM, and Locuslink (www.ncbi.nlm.nih.gov/LocusLink) web sites, we automatically extracted the gene name symbols and disease mentions from all Medline article titles and abstracts. Where possible, we replaced alias symbols with official ones. We also extracted key words related to a certain disease and used them to determine which genes were statistically correlated with this disease.

To test a gene for statistical relevance to a disease, we simply compared the observed number of gene–disease co-occurrences to the number we would expect given no correlation. Because the distribution of co-occurrences of two uncorrelated terms follows a binomial distribution, a value of observed gene–disease co-occurrences more than one SD greater than the binomial expected value indicates correlation. This statistical method is preferable to the "term frequency, inverse document frequency" metric, because it accurately handles infrequently mentioned genes, which are very common.

The final step in obtaining data was to remove false positives, which occur frequently because gene symbols generally coincide with other abbreviations having nothing to do with genes. For example, the symbol HDC, representing the gene histidine decarboxylase, was commonly used in the literature as an abbreviation for high dose chemotherapy. We disambiguated the data, using a method shown in ref. 2, which yielded unambiguous symbol identifications with a low error rate.

Gene Graph

The creation of gene graphs from the co-occurrence data was performed following a well known procedure (21, 23). Each vertex in the graph represents a gene, and an edge exists between two vertices if the genes they represent co-occur at least once. We did not use weighted edges. In creating the graph, we neglected articles published before 1990 and articles that listed more than five genes, as mentioned in the Introduction.

The resulting graph has a power law distribution in its degree. That is, the number of vertices of degree x is given by $Ax^{-\beta}$, where $\beta < 0$. This is shown in Fig. 1, where we plot the data on a log–log scale for gene graphs corresponding to several diseases.

The properties of such power law graphs have been extensively studied (27–29). It has been shown that random graphs with

[h]This number was determined by comparing the International Standard Serial Numbers of the journals in the Medline listings of the 7,985 abstracts involved in creating the network of genes related to colon cancer.

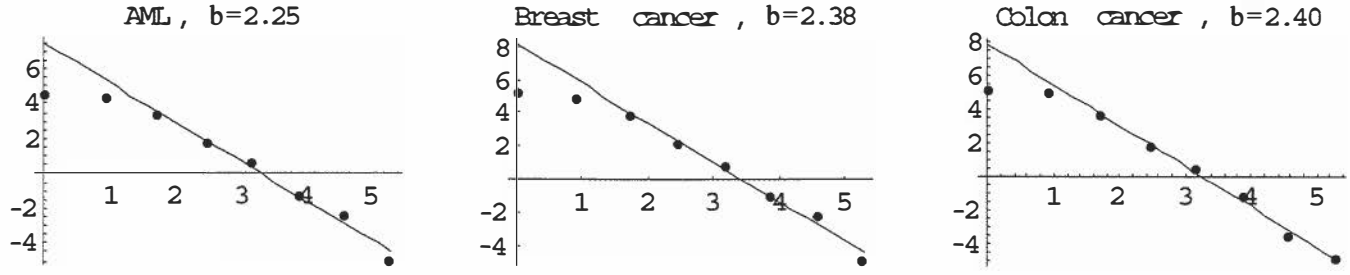


Fig. 1. The number of vertices (*y* axis) is plotted against the degree of the vertex (*x* axis) for several diseases on a log–log scale. We followed the usual binning procedure in plotting the data. The deviation from the power law for low vertex degree is typical. AML, acute myelogenous leukemia.

$2<\beta<3.5$ consist of one giant connected component and other small components of size $O(\ln(N))$ (28). Here N is the size of the graph, and β is the power law exponent. The component structure of the gene graphs agrees with the predictions of ref. 28 for random graphs, as shown in Table 1.

Because the smaller components contain few genes with few neighbors, they are of limited interest. They usually consist of little-known genes that have not been related to other genes. In what follows, we focus exclusively on identifying communities within the giant component.

Partitioning the Graph into Communities

There is no formal definition for a community of vertices within a graph. A graph can be said to have community structure if it consists of subsets of genes, with many edges connecting vertices of the same subset but few edges lying between subsets (4). Finding communities within a graph is an efficient way to identify groups of related vertices.

As mentioned in the Introduction, the community discovery process we use is based on that of GN (the GN process or method), which has been shown to identify communities in graphs with known community structure to a high degree of accuracy (4). Our modifications were necessary to make the method applicable to gene graphs, which are large and are created from source information that may by nature be incomplete or flawed. In particular, we identify many possible community structures and average them into a final list of communities. The statistical character of this step provides a more accurate picture of the complicated nature of community structure of a gene graph, without undermining the effectiveness of the basic principle of the algorithm.

Table 1. Sizes of connected components in several gene graphs

	Components	
Disease (no. of statistically relevant genes)	Size	No.
Acute myelogenous leukemia (488)	460	1
	4	1
	3	4
	2	6
Breast cancer (816)	686	1
	6	2
	5	1
	4	5
	3	9
	3	33
Colon cancer (682)	561	1
	4	4
	3	15
	2	30

A concept central to the community discovery process is the betweenness centrality (hereafter betweenness) of a vertex or edge. The betweenness of an edge AB (or a vertex A) is defined as the number of shortest paths between pairs of other vertices that contain AB (or A). As mentioned before, this concept was introduced (3) as a measure of influence of an individual, with respect to information flow, within a social network., However, it was noticed (4) that betweenness may also be used to identify communities within a graph, because intercommunity edges (those that lie between different communities) are much more likely to have a higher betweenness than intracommunity edges (edges that lie within one community).

To explain the community discovery process, we consider as a first example the small graph shown in Fig. 2. This graph consists of two well defined communities: the four vertices denoted by squares, including vertex A, and the nine denoted by circles, including vertex B.

In the graph of Fig. 2, edge AB has the highest betweenness. If we were to remove it, the graph would split into two connected components, the square and circle communities. This illustrates the idea behind the GN method of imposing community structure on a graph. One repeatedly identifies intercommunity edges by the criterion that they have higher betweenness than intracommunity edges and removes them. This procedure splits the giant component into many separate components, which coincide with the communities of the original graph.

It is important to note that the removal of an edge strongly affects the betweenness of many others, so that one must repeatedly recalculate the betweenness of all edges. To do this quickly, we used the fast algorithm of ref. 29 or 30.

At a certain point in our procedure, as opposed to the GN method, we stop removing edges from a component when we cannot further meaningfully subdivide it into communities; for example, as in Fig. 2, after removing edge AB. This allows us to obtain distinct communities of nodes, such as the circles and squares of Fig. 2. What criterion tells us when to stop?

Structurally, a component of five or fewer vertices cannot consist of two viable communities. The smallest possible such component is size 6, consisting of two triangles linked by one edge (Fig. 3).

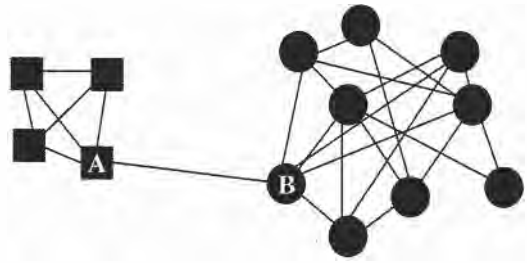


Fig. 2. A graph consisting of two communities.

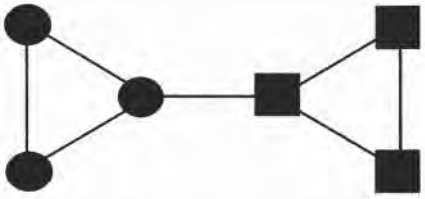


Fig. 3. The smallest possible graph consisting of two communities.

Components of size ≥6 can also be individual communities, like the group of nine in Fig. 2. The criterion we used to identify this type of component as a community was that the largest betweenness of any edge in the component did not exceed $N - 1$, where N is the number of vertices in the component.

This threshold is based on the betweenness of an edge connecting a leaf vertex, or vertex of degree one, to the rest of the graph. Consider the graph of Fig. 4 below. It is clear that it consists of just one community. Applying the Brandes algorithm, we find that edge XY has the highest betweenness, indicating that the size of the largest distinct community within the graph has size 1. That is, there are no distinct communities within the graph. In general, the single edge connecting a leaf vertex (such as X in Fig. 4) to the rest of a component of N vertices has a betweenness of $N - 1$, because it contains the shortest path from X to all $N - 1$ other vertices. If no edge's betweenness exceeds $N - 1$, therefore, we can identify the component as a community.[i]

We can now explain the need to neglect survey-type articles that list many genes in creating our graph. The genes listed in these articles will all be linked to one another, forming a complete subgraph K_n. Such a grouping is very tightly knit and will likely not be split into different communities. This situation, due only to the survey article, may not accurately reflect the interactions between the genes. It is possible that a few articles mention many genes that are in fact functionally related, but in this case it is likely that the genes will be linked by other articles that discuss them three or four at a time.

Communities Consist of Functionally Related Genes. The communities thus created consist of genes that were strongly interrelated in the literature. Most, but not all, gene co-occurrences imply a functional relation; genes may also co-occur in an article abstract because of physical proximity, similarity of nomenclature or structure, historical association, or other reasons. However, because such nonfunctional edges are a minority, they are highly likely to be intercommunity, because the neighbors of two nonfunctionally related genes are unlikely to be linked.

For example, genes *S100A4* and *S100A6* are members of the S100 family and co-occur twice in articles related to colon cancer, but they are not functionally related (Medline PMIDs 10389988 and 10952782). In our results, *S100A4* and *S100A6* do not occur in a community together. The neighbors of one are not linked to the neighbors of the other, which causes them to be placed in separate communities. Further examples are given in *Results*.

[i]It is not in general true that an intercommunity edge must have betweenness greater than $N-1$, although such a situation is extremely unlikely in a power law graph. For a community of size m within a graph of size N, there is a total betweenness of $m(N-m)$ divided among the edges connecting the community to the graph. So, if there are more than m such edges, it is possible that none of them will have betweenness greater than N. However, remember that few of these edges, or the extracommunity vertices they connect, should be adjacent, because otherwise m would not be a community. Even in GN's highly nonpower law college football graph, the criterion only occasionally fails when an intercommunity edge has a betweenness slightly less than $N-1$.

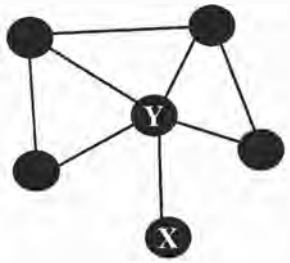


Fig. 4. A partitioning algorithm should not separate this graph into two communities.

Multiple Community Structures. The process of assigning the nodes of a graph to communities may be called identifying a community structure on the graph. In the small examples given thus far in Figs. 2–4, there was only one reasonable community structure on each graph, because each node clearly belonged to only one community. In contrast, complex real-world graphs contain many "ambiguous" nodes that can be said to belong to two or more different communities due to their placement in the graph. An example, described in detail later, is the subgraph in Fig. 5, in which node B is ambiguous. Gene graphs include many ambiguous genes that belong to several communities, both in the context of the graph and in the context of biological function.

Therefore, if we identify only one community structure on a real-world graph, such as a gene graph, we could only hope to be somewhat accurate in classifying the nodes. A large amount of information concerning ambiguous genes and communities related through ambiguous genes would be lost.

Our resolution to this problem is to identify many plausible community structures on the graph and compare them. To do this, we make a modification to the GN process that introduces an element of randomness into which edges of very high betweenness are removed early in the process. Tightly knit communities are not affected by the order of edge removal and will eventually be identified no matter which high-betweenness edges are removed first. However, the eventual placement of ambiguous genes is strongly affected by which high-betweenness edges are removed first. By varying which high-betweenness edges are removed early in the process, we may therefore identify many community structures on a graph. By then comparing the structures, we can easily identify tightly knit communities, which do not vary from structure to structure, and ambiguous genes, which migrate from group to group.[j]

The subgraph of Fig. 5 illustrates why the order of edge removal affects the placement of ambiguous genes and the need for multiple community structures. This subgraph consists of two communities, one on the left including vertex A and another on the right including C. Among its edges, BC initially has the highest betweenness, and AB's betweenness is also high. Once we remove BC, however, AB becomes an intracommunity edge with low betweenness, and it will never be removed. Gene B will eventually be placed in a community with gene A. Had we removed AB first, BC would be rendered intracommunity, and gene B would end up in the community with C. Moreover, in considering Fig. 5, it is not clear where B should end up. B is

[j]This process is essentially a form of soft clustering, although it differs significantly from existing methods of soft clustering. These methods (see ref. 34 for an example) are essentially restricted to clustering objects such as documents that comprise many individual elements (e.g., words). The words of one document are compared to the words of another, and a relative closeness can be established. The soft clustering presented here is affected only by a node's placement in the graph, not by a comparison of elements comprising neighboring nodes.

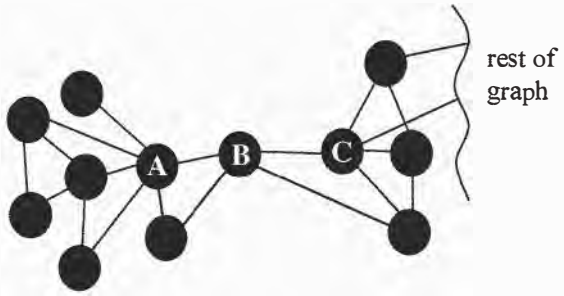


Fig. 5. In this graph, it is unclear to which community node B belongs.

ambiguous and could rightfully be considered to be a part of both communities. If we considered multiple community structures on this subgraph, we would see that B ended up in the community with A in some structures (the ones in which BC was removed early) and in the community with C in others (those where AB was removed). By comparing the structures, we could see that B really plays a role in both communities and (depending on the meaning of the links in the graph) possibly ties the communities together.

To describe in detail our method of identifying multiple community structures, we briefly describe how the Brandes algorithm (31) computes the betweenness for all edges in a graph, and how the GN process decides which edge to remove at each step. We then explain our modification, which allows for the identification of multiple community structures.

The first step of the Brandes algorithm is to find the shortest paths from all vertices to one "center" vertex using a breadth first search. In the second step, the contributions of these paths to the betweenness of each vertex or edge are added to a running total. The center is then switched and the above steps repeated. After every vertex in the component has been the center, we will have considered every shortest path twice, and the running totals for each vertex or edge will equal twice the betweenness of that vertex or edge. At this point in the GN procedure, one simply chooses the edge of highest betweenness and removes it. However, this choice is somewhat arbitrary, because there are likely to be many intercommunity edges in the graph.

Our modification is as follows. Instead of using every vertex as the "center" once in the Brandes algorithm, we cycle randomly through at least m centers (where m is some cutoff[k]) until the betweenness of at least one edge exceeds a threshold, again based on the betweenness of a "leaf" vertex.[l] We then remove the edge whose betweenness is highest at that point and repeat. Because the betweenness of the edge we remove exceeds the threshold, it is very likely to be intercommunity. We continue removing edges from each component in this way, until all of the components become small.[m] We then perform the full Brandes algorithm and remove the edge of highest betweenness from each component until it is resolved into communities.

The random nature of the modification allows us to change the order in which edges are removed, because the edge with highest betweenness after m centers have been considered will vary depending on which centers are considered. Our modified process can therefore be applied repeatedly to identify different plausible community structures on the graph.

This process may erroneously remove an intracommunity edges, which can happen if a large percentage of the centers considered lies in one community. In a large graph with many small communities, this probability is small, especially because we perform only the modified removal step in large components. Additionally, when we compare many different community structures, anomalous placements due to errors will be suppressed.

Applying this modified process n times, we obtain n community structures imposed on the graph. We can then compare the different structures and identify communities, as well as the strength of each gene within the community. For example, after imposing 45 structures on our graph, we might find: a community of genes A, B, C, and D in 20 of the 45 structures; a community of genes A, B, C, D, and E in another 20; and one of genes A, B, C, D, E, and F in the remaining 5. We report this result in the following way: A(45) B(45) C(45) D(45) E(25) F(5), which signifies that A, B, C, and D form a well defined community, E is related to this community but also to some other(s), and F is only slightly, possibly erroneously, related to it.

Aggregating Communities from Different Structures. To aggregate communities from different structures and obtain a final list of communities in the form {A(45) B(45) C(45) D(45) E(25) F(5)}, we use a procedure that is straightforward but rather tedious to explain because of the terminology. To summarize briefly, we create an initial "master list" M^1 of communities by choosing one structure at random from among our set of N (in our experiments, $n = 50$). We then perform $N - 1$ steps, each consisting of comparing one of the remaining $N - 1$ structures S to the master list, and based on the results of the comparison, aggregating S into the list. The final master list, obtained by aggregating all N structures, is the final result of the entire algorithm.

Let us introduce the notation M^t to denote the state of the master list created from aggregating the first t structures (chosen in arbitrary order from the set of structures we found). At step $t + 1$, we select a structure S from among those we have not yet considered and compare its communities to the communities of M^t. S is aggregated into M^t based on the results of the comparison, creating an updated master list M^{t+1}.

M^t is a list of communities, and we will denote the kth community of M^t by M^t_k. The numbering system of communities in the list is arbitrary and serves only to distinguish them. Each community M^t_k of M^t is a collection of genes and associated weights $\{(\beta_j, \rho_j)\}$. The weight ρ_j associated with gene β_j in a community M^t_k indicates how strongly it "belongs" to M^t_k. To be precise, ρ_j is the number of structures, out of the t we have aggregated to form M^t, in which β_j has been associated with the community that evolved (because structures were aggregated) into M^t_k. This will be clearer when we explain the aggregation step. The communities evolve very little as the structures are aggregated, because the structures are on the whole quite similar. Thus the weight as defined is an accurate indication of how strongly a gene belonged to a community. One might expect that, because the communities in the master list evolve, the final result would depend on the order of aggregation. To the contrary, we found the order of aggregation had little effect on the final result due to the similarity of the different structures.

The details of the matching of communities S to the communities of M^{t-1}, and of the aggregation of S into M^{t-1} to form M^t, are as follows. A basic metric to compare two communities A and

[k] The cutoff we used was $m(N) = 10\log(N) - 25$, where N is the size of the component. This function has $m(50) \approx 15$, and $m(800) \approx 41$. We found that 15 was a reasonable number of centers to consider for a component of size 50, whereas 40 centers is more than enough for any component, however large. Basically, an intracommunity edge will be erroneously removed if we repeatedly choose centers from the same community. For a component of 50 vertices and 4 communities, the probability of choosing 8 of 15 centers from one community is ≈1%. For a large component with many communities, the probability of error is very low for a cutoff of 40 centers.

[l] The value of the threshold in this case is $(N + 1)/2 - 1$, where N is the size of the component, and i is the number of centers that have been considered up to that point in the process.

[m] We never attempted to precisely define "small." We used values in the 35–50 node range and, as one might expect, it made little or no difference in the final result. An exact definition would depend on the community size, the graph size, and a desired probability of error (see discussion on this page). However, even when we used a number as large as 50, the randomness of the method was sufficient to produce a slightly different community structure every time.

B of genes $\alpha_1, \alpha_2, \ldots, \alpha_n$, and $\beta_1, \beta_2, \ldots, \beta_m$, respectively, the traditional union/intersection metric

$$d(\mathrm{A}, \mathrm{B}) = \frac{\mathrm{A} \cup \mathrm{B}}{\mathrm{A} \cap \mathrm{B}} = \frac{\sum_{i,j} \delta_{\alpha_i, \beta_j}}{n + m - \sum_{i,j} \delta_{\alpha_i, \beta_j}}$$

Here n and m are the number of genes in communities A and B, so the sums run over $i = 1,\ldots,n$ and $j = 1,\ldots,m$. This notation is overcomplicated but useful for comparison to the weighted metric below. The sum over Δ functions just means we are counting how many of the genes in A and B are the same. The metric $d(\mathrm{A}, \mathrm{B})$ has a value between 0 and 1 and will be larger for a closer match. To compare a community A of S to a weighted community M_k^{t-1} of M^{t-1}, we modify the traditional union/intersection metric to include the weight:

$$d_M(\mathrm{A}, M_k^{t-1}) = \frac{\sum_{i,j} \frac{\rho_j}{t-1} \delta_{\alpha_i, \beta_j}}{n + \sum_j \frac{\rho_j}{t-1} - \sum_{i,j} \frac{\rho_j}{t-1} \delta_{\alpha_i, \beta_j}}$$

By comparing each community in one structure to all of the communities in the other using the weighted metric, we can find the closest match for each one.

Once a closest match for each community of S is found from among the M^{t-1}, the communities of S are aggregated into M^{t-1}. If community A of S is matched to community M_k^{t-1}, we combine A and M_k^{t-1} by incrementing the weights of the genes common to A and M_k^{t-1} and appending the genes in A that were not in M_k^{t-1}. For example, suppose that community {C, D, F} in S is matched to {B(5)C(5)D(3)E(5)} in M^{t-1}. We would update this community to become {B(5)C(6)D(4)E(5)F(1)} in M^t; that is, C and D would be incremented, and F would be appended.

Occasionally, two or more communities in the structure were matched to one in M and vice versa. In this case, we assumed that the intracommunity edge had been erroneously removed to divide one community into two or more, either in the structure or the master list (in that case, it would have been in one of the previous structures aggregated into the master list). We thus melded the divided communities into one, altering M if need be, and then updated M as described above. This step could create a problem if one ended up with huge communities at the end, but we found that in general the largest communities in the final result had only 10 or 15 more genes than the largest communities in each individual structure, which incidentally indicates that our edge removal algorithm had a low error rate.

The entire process of determining community structure is displayed in Table 2.

Table 2. Algorithm for determining community structure

A. For *n* iterations, repeat {
1. Break the graph into connected components.
2. For each component, check to see whether component is a community.
   a. If so, remove it from the graph and output it.
   b. If not, remove edges of highest betweenness, using the modified Brandes algorithm for large components and the normal algorithm for small ones. Continue removing edges until the community splits in two.
3. Repeat step 2 until all vertices have been removed from the graph in communities.

}

B. Aggregate the *i* structures into a final list of communities.

Results

We applied the above technique using key words related to colon cancer. We considered articles that mentioned at least one of colon, colorectal, colonic, or gastrointestinal, and at least one of cancer or carcinoma. We identified 682 genes that were statistically correlated with colon cancer and that co-occurred in these articles with at least one other correlated gene. The graph of this co-occurrence network consisted of a giant component of 561 genes and other uninteresting smaller components (Table 1). The community discovery algorithm split the giant component into 79 different communities, with sizes ranging from 2 to 50 genes.

To present the usefulness of our results, we discuss features of these communities that demonstrate the utility of our method. Used in conjunction with the Medline and OMIM web sites, these communities allow us to suggest undocumented connections between genes of one community and between genes in different communities. They also demonstrate that our method tends to separate genes that co-occurred but were not functionally related into different communities, as discussed in *Gene Graph*. Genes that occur in two or more communities can indicate a link between the genes of each community.

We have published a full list of communities related to colon cancer and other diseases on our web site. Here we simply present one community to demonstrate the format of the results, discuss its features, and briefly mention similar features of other communities.

Table 3 shows one community of genes related to colon cancer from our results. Genes in this community are related to the overexpression of prostaglandin-endoperoxide synthase 2 (*PTGS2*), in colon cancer. Although *PTGS2* is the official HUGO symbol, this gene is very commonly called cyclooxygenase 2 (*COX-2*), and we will use this term.

The features of this community suggest the following possibilities: connections between some of the genes that co-occur with *COX-2*, but not each other; good reasons why many of the neighbors of *COX-2* are not in this community; and possible connections to other communities via progesterone E synthase (*PGES*) and lymphoid enhancer-binding factor 1 (*LEF1*). We investigated these possibilities and present the results below.

Implied Connections. This community suggests a possible connection between the phospholipase A2 genes in this group and the gene *FACL4*. A Medline search for *FACL4* or its alias *ACS4* with each of *PLA2*, *SPLA2*, *PLA2G4*, and *PLA2G2A* turned up no result, and the OMIM entry for *FACL4* has no mention of phospholipase A2. Nevertheless, by examining the abstracts of articles in which these genes were found, we see that these genes are related by their function, via *COX-2* and arachidonic acid. COX enzymes convert arachidonic acid to prostaglandins (Medline PMID 11274413, for example). The three phospholipase A2 genes in the group {*SPLA2*, *PLA2G4* [also known as *cPLA* (2)], *PLA2GA2*} are all sources of arachidonic acid (PMID 10706128, for example) and are thus related to *COX-2*. However, we found that the *FACL4* enzyme also uses arachidonic acid, and that "the cellular level of unesterified arachidonic acid is a general mechanism by which apoptosis is regulated and that *COX-2* and *FACL4* promote carcinogenesis by lowering this level" (PMID 11005842). This indicates a clear link between the phospholipase A2 family of genes and *FACL4* in carcinogenesis. It would have been time consuming for a researcher to ascertain this connection manually from Medline; even a search for arachidonic acid and colon cancer together produces 119 abstracts to sift through. Additionally, during this brief literature search, we discovered that nonsteroidal antiinflammatory drugs (NSAIDs) function by suppressing *cPLA2* (*PLA2G4*) mRNA expression and thus depriving COX-2 of arachidonic acid.[k] Our method therefore suggests that these drugs may possibly affect

Table 3. A sample community of nine genes from our results for colon cancer

Gene symbol	Weight in community	Overall mentions with colon cancer	Neighbors with colon cancer
PTGS2	50	263	*PTGS1* DLD* MLH1* BCL2* PLA2G2A PLA2G4 APC* ERBB2* PGES ERBB3* PLA2 ACL4 WNT1* GRP* GRPR* LEF DLR* TCF4* TCF* MYB* VEGF* NOS2A TP53* MADH4* EGFR* S11* PDCD4 BRCA1* BRCA2* MSH2* ERBB4*
PLA2G2A	50	12	*APC* PTGS2 PLA2G4 TP53* NF2* DCC* MLH1* SPLA2*
PLA2G4	50	1	*PLA2G2A PTGS2*
SPLA2	50	4	*PTGS2 PLA2G2A*
FACL4	50	1	*PTGS2*
NOS2A	50	7	*PTGS2*
PDCD4	50	1	*PTGS2*
PGES	18	2	*ERBB2* PTGS2 ERBB3**
LEF1	5	18	*WNT1* TCF* PTGS2 TCF4* APC* FRA1* PLAUR* MYC* MMP7* TCF7**

Here score in community denotes the number of community structures, out of 50, in which each gene was placed in this community (*Partitioning the Graph into Communities*). Genes with a score of 50 were members of this community only; genes with a lower score were members of this community and others.
*Neighbor not in community.

FACL4 expression, although a Medline search of NSAID and *FACL4* turned up no results.

Absent Neighbors. In examining neighbors of *PTGS2* (*COX-2*) not present in this community, we noticed in particular the similarly named gene *PTGS1* (also known as *COX-1*). These two genes are isoforms of cyclooxygenases (PMID 9099957, for example); they co-occurred in 70 articles related to colon cancer and 1,500 articles overall. However, they have been shown to regulate colon carcinoma-induced angiogenesis by two different mechanisms (Medline PMID 9630216). *COX-2* has also been shown to be expressed much more frequently than *COX-1* in tumors and less frequently in normal tissue (PMID 7780968, for example; note the use of the alias *PGHS-1* or -2 for *COX-1* or *-2* in this article) The separation of *COX-1* and *-2* into different communities thus accurately reflects our current knowledge about how these genes function in relation to colon cancer. Although the enzymes they code for are structurally very similar, *COX-2* plays a strong role in colon cancer, whereas *COX-1*'s role is weaker and by a different mechanism.

Several other neighbors of *PTGS2*, such as *MLH1*, *BRCA1*, *BRCA2*, and *MSH2*, also proved to be weakly or nonfunctionally related. However, a few of *PTGS2*'s noncommunity neighbors have been tentatively identified as functionally related, such as *GRP* and *GRPR* (*GRP receptor*; PMID 11292836) and *EGFR* (PMID 9012840). For this reason, we include a list of all neighbors of each gene in the results as a secondary list of possible connections to explore.

Links to Other Communities. We also looked for links to other communities through the genes *PGES* and *LEF1*, both of which show a weak connection to the COX-2 community and were often placed in other communities.

Both searches yielded good results. *PGES* co-occurs with other genes only once, in an abstract with *COX-2*, *ERBB2*, and *ERBB3*. Examining this abstract, we find a link between the *COX-2* pathway and autocrine/panacrine activation of *HER2/HER3* (also known as *ERBB2* and *ERBB3*; 9927187). The *ERBB* genes are present in another community of 25 genes. In conjunction with the previous discussions about arachidonic acid, there is a possible link between not only *COX-2* but all of the genes related to arachidonic acid (most of which never co-occur with *ERBB2* or *-3*) to any gene related to the autocrine/panacrine activation of *ERBB2/ERBB3*. This conclusion depends on knowledge of many articles, in particular PMID 9927187, and could easily escape notice in a manual search.

LEF1 was found with *COX-2* in only one article (PMID 10834941). It states that "NO (nitric oxide) may be involved in *PGHS-2* (*COX-2*) overexpression in conditionally immortalized mouse colonic epithelial cells. Although the molecular mechanism of the link is still under investigation, this effect of NO appears directly or indirectly to be a result of the increase in free soluble β-catenin and the formation of nuclear *β-catenin/LEF-1* DNA complex." This article indicates a possible connection between COX-2, NOS2A (nitric oxide synthase, responsible for the production of NO) and the very important colon cancer gene *β-catenin*.

Importance of Alias Symbols. As a last note, this community demonstrates the crucial importance of considering alias symbols when extracting gene names. The aliases *COX-2*, *PGHS-2*, *NOX2*, and *cPLA (2)* were very commonly used in articles that tied this community together

Other Results. Here we present similar results from two other communities: A connection between *PXR* (pregnane X receptor) and *GP170* (P-glycoprotein) is indicated because they are placed together in a community. *PXR* is implicated in the induction of the *MDR1* gene (PMID 11297522), whereas *MDR1* expression has been associated with the expression of functional P-glycoprotein (PMID 10334913). A Medline search turns up no results for *GP170* or *GP-170* with *PXR* or its aliases *PAR*, *SXR*, and *NR1i2*.

Another probable undocumented connection between *GP200-MR6* and *STAT6*, via *IL-4* and its receptor *IL-4R* is suggested by their placement together in a community. *IL-4* induces *STAT6*, which is involved in mediating activation of *IL-4R* gene expression (PMID 8810328), whereas *GP200-MR6* has been shown to be functionally associated with *IL-4R* (PMID 9178815). This example demonstrates the power of an automated method to bring together information from disparate, old sources (cited articles from *J. Biol. Chem.*, Oct 11, 1996 and *Int. J. Cancer*, May 16, 1997).

Although large communities are more difficult to analyze for the nonexpert, we were nevertheless able to draw some conclusions. For example, we considered a 30-gene community largely concerned with apoptosis and genes related to *BCL-2*, containing in particular the gene *TRAIL*. *TRAIL* has been shown to induce procaspase-8 activation, triggering caspase-dependent apoptosis in colon cancer cells (PMID 11245478). It could thus be related to the function of genes such as *BCLX*, *BCLXS*, etc., which we find in this community but which do not co-occur with *TRAIL* via the genes *BCL-2* and *CASP8*.

A good example of nonfunctionally related genes with similar names that are placed in different communities is *MMP11* and *MMP9* (PMID 8645587). Often nonfunctionally related neighboring genes do appear together in one community in a small

number of structures (see *Partitioning the Graph into Communities*) but appear in different communities in the majority of structures. Examples of this include *CYP3A4* or *CYP3A5* and *CYP1A2* (PMID 9202751) as well as *SMAD3* and *SMAD5* (PMID 10446110 and 11196171, for example; *SMAD2* and *SMAD4* are aliases for *MADH2* and *MADH4*, respectively).

Conclusion

We have presented a data-mining technique for biological literature that produces detailed results while extracting only very simple data from each article abstract and title. The method produces a list of communities of functionally related genes that are designed to summarize available information and indicate genes that are likely to be complementary in their function. The genes within a community are weighted, indicating how strongly they belong to the community. We show that the communities produced in the case of colon cancer have interesting features that give one insight into the function of the component genes.

The identification of many similar community structures on each gene graph allows us to recognize those genes that belong in two or more different communities. In this sense, our method produces a richer result than previous methods that impose one rigid structure on the graph. This idea could be applied to social and other networks where individuals play a role in more than one community.

We introduce two statistical components into the process, which lessen the inevitable errors of text mining in the biological literature, particularly severe in our case because of the complex young nomenclature system for genes. However, our method retains the ability to detect relations among rarely mentioned genes, one of its strongest features.

To reiterate an important point from the Introduction, our results are not meant to perfectly model biological reality, only to function as a tool for biologists. It was not possible to compare our communities to a database or list of groups of related human genes, because such a list does not exist. The only justification we can provide that our communities were "accurate" is to cite ref. 4, in which the GN method was shown to be very effective in identifying communities. In fact, because genes within a community are linked by edges from a co-occurrence, it is almost certain that they are related somehow. A much more interesting measure of the effectiveness of the method is whether it separates genes that should be separated.

The factor that most limits our results is the absence of many gene symbols from HUGO and other online databases. Hopefully, these databases will soon be more complete. Related problems are the unorganized nomenclature system for human genes (see discussion in ref. 31) and small modifications to recognized symbols introduced by many authors, such as the addition of hyphens, parentheses, or spaces, which make the symbols difficult to detect. Efforts are being made to standardize the gene nomenclature system (33).

A less acute limiting factor was the placement of many genes in either large and very small communities in our results. Although still a step forward from raw co-occurrence data, such communities are of limited usefulness; they often did not provide much insight into the function of their component genes, other than that the genes were rarely related to others in the context of colon cancer. If such genes were more commonly mentioned in other contexts, a search using other diseases or key words would likely turn up more interesting communities with these genes. Large communities were difficult for us to analyze but nevertheless yielded some interesting results. These communities contained many of the most commonly mentioned genes in connection with colon cancer, such as APC and TP53. Strangely, a search for colon cancer genes is probably not the most efficient way to study these genes, which are simply too highly linked in this context. Instead, one could perform other searches with other key words, hoping to focus on particular aspects of these genes' function by confining them to smaller more informative communities.

We believe that large communities are a product of graph topology, not of the threshold we use to stop subdividing a community or of the aggregation process. To further subdivide large communities, one could consider a weighted graph, where the weight corresponds to the (normalized) number of times the two genes co-occur. This could increase the "distance" between, for example, two commonly studied distantly related co-occurring genes. They would then not end up in the same community and, more importantly, would not glue a false community together. The simplest such weighting would be to neglect all links below some (normalized) threshold weight. Another resolution to the problem of large communities would be to refine the step that aggregates the community structures into one result.

We thank Lada Adamic, Eytan Adar, and Melissa Wilkinson for many useful discussions.

1. Adamic, L., Wilkinson, D., Huberman, B. & Adar, E. (2002) in *Proceedings of the IEEE Bioinformatics Conference* (Institute of Electrical and Electronic Engineers, Los Alamitos, CA), pp. 109–117.
2. Adar, E. (2004) *Bioinformatics*, in press.
3. Freeman, L. (1977) *Sociometry* **40,** 35–41.
4. Girvan, M. & Newman, M. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 8271–8276.
5. Wu, F. & Huberman, B. A. (2004) *Eur. J. Phys. B*, in press.
6. Tyler, J. R., Wilkinson, D. M. & Huberman, B. A. (2003) *Proceedings of the International Conference on Communities and Technologies* (Kluwer, Amsterdam).
7. Blaschke, C., Andrade, M., Ouzounis, C. & Valencia, A. (1999) in *Proceedings of the AAAI Conference on Intelligent Systems in Molecular Biology* (American Association for Artificial Intelligence Press, Heidelberg), pp. 60–67.
8. Ng, S. K. & Wong, M. (1999) *Genome Inf.* **10,** 104–112.
9. Humphreys, K., Demetriou, G. & Gaizauskas, R. (2000) *Pac. Symp. Biocomput.* **5,** 502–513.
10. Rindflesch, T. C., Tanabe, L., Weinstein, J. N. & Hunter, L. (2000) *Pac. Symp. Biocomput.* **5,** 514–525.
11. Thomas, J., Milward, D., Ouzounis, C., Pulman, S. & Carroll, M. (2000) *Pac. Symp. Biocomput.* **5,** 538–549.
12. Friedman, C., Kra, P., Yu, H., Krauthammer, M. & Rzhetsky, A. (2001) *Bioinformatics* **17,** S74–S82.
13. Pustejovsky, J., Castaño, J., Zhang, J., Cochran, B. & Kotecki, M. (2002) *Pac. Symp. Biocomput.* **7,** 362–272.
14. Tanabe, L., Scheft, U., Smith, L., Lee, J., Hunter, L. & Weinstein, J. (1999) *BioTechniques* **27,** 1210–1217.
15. Craven, M. & Kumlien, J. (1999) in *Proceedings of the ISMB Conference* (International Society for Computational Biology, Brisbane, Australia), pp. 77–86
16. Shatkay, H. & Wilbur, W. (2000) in *Proceedings of the IEEE Conference on Advances in Digital Research* (Institute of Electrical and Electronic Engineers, Los Alamitos, CA), pp. 183–192
17. Raychaudhuri, S., Chang, J., Sutphin, P. & Altman, R. (2002) *Genome Res.* **12,** 203–214.
18. Andrade, M. & Valencia, A. (1997) in *Intelligent Systems for Molecular Biology* (American Association for Artificial Intelligence Press, Heidelberg), pp. 25–32.
19. Fukuda, K, Tsunoda, T., Tamura, A. & Takagi, T. (1998) *Pac. Symp. Biocomput.* **3,** 705–716.
20. Proux, D., Rechenmann., F., Julliard, L., Pillet, V. & Jacq, B. (1998) in *Genome Informatics Workshop* (Universal Academic, Tokyo), pp. 72–80.
21. Jenssen, T.-K., Laegrid, A., Komorowski, J. & Hovig, E. (2001) *Nat. Genet.* **28,** 21–28.
22. Stephens, M., Palakal, M., Mukhopadhyay, S. & Raje, R. (2001) *Pac. Symp. Biocomput.* **6,** 483–396.
23. Stapley, B. & Benoit, G. (2000) *Pac. Symp. Biocomput.* **5,** 529–540.
24. Masys, D., Welsh, J., Fink, L., Gribskov, M., Klacansky, I. & Corbeil, J. (2001) *Bioinformatics* **17,** 319–326.
25. Shatkay, H., Edwards, S. & Boguski, M. (2002) in *IEEE Intelligent Systems, Special Issue on Intelligent Systems in Biology* (Institute of Electrical and Electronic Engineers, Los Alamitos, CA), pp. 45–53.
26. Raychaudhuri, S., Schutze, H. & Altman, R. (2002) *Genome Res.* **12,** 1582–1590.
27. Huberman, B. A. & Adamic, L. (1999) *Nature* **401,** 131.
28. Aiello, W., Chung, F. & Lu, L. (2000) in *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York), pp. 171–180.
29. Albert, R. & Barabasi, A.-L. (2002) *Rev. Mod. Phys.* **74,** 47–97.
30. Newman, M. (2001) *Phys. Rev. E* **64,** 026118-1–026118-17.
31. Brandes, U. (2001) *J. Math. Soc.* **25,** 163–177.
32. Pearson, H., (2001) *Nature* **411,** 631–632.
33. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S. & Eppig, J. (2000) *Nat. Genet.* **25,** 25–29.
34. Tishby, N., Pereira, F. & Bialek, W. (1999) in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing* (UIUC Press, Champaign–Urbana, IL), pp. 368–377.

Tracking evolving communities in large linked networks

John Hopcroft*, Omar Khan†, Brian Kulis‡, and Bart Selman*§

*Department of Computer Science, Cornell University, Ithaca, NY 14853; †Google, Inc., Mountain View, CA 94043; and ‡Department of Computer Science, University of Texas, Austin, TX 78712

We are interested in tracking changes in large-scale data by periodically creating an agglomerative clustering and examining the evolution of clusters (communities) over time. We examine a large real-world data set: the NEC CiteSeer database, a linked network of >250,000 papers. Tracking changes over time requires a clustering algorithm that produces clusters stable under small perturbations of the input data. However, small perturbations of the CiteSeer data lead to significant changes to most of the clusters. One reason for this is that the order in which papers within communities are combined is somewhat arbitrary. However, certain subsets of papers, called natural communities, correspond to real structure in the CiteSeer database and thus appear in any clustering. By identifying the subset of clusters that remain stable under multiple clustering runs, we get the set of natural communities that we can track over time. We demonstrate that such natural communities allow us to identify emerging communities and track temporal changes in the underlying structure of our network data.

Emergent properties of large linked networks have recently become the focus of intense study. This research is driven by the increasing complexity and importance of large networks, such as the World Wide Web, the electricity grid, and large social networks that capture relationships between individuals. Real-world networks generally exhibit properties that lie somewhere in-between those of highly structured networks and purely random ones (1–4). So far, most research has focused on using static properties, such as the connectivity of the nodes in the network and the average distance between two nodes, to explain the complex structure. However, these networks generally evolve over time and so temporal characteristics are a key source of interest. Our goal in this paper is to provide techniques for the study of the evolution of large linked networks.

In our approach, we use agglomerative clusterings of the linked network. By clustering the network at different points in time, we study its temporal evolution. This approach places a new burden on the underlying clustering method. Clustering methods can be surprisingly sensitive to minor changes of the input data. For obtaining a static view of the higher-level structure of the data, such instabilities may be acceptable because the resulting hierarchy often already reveals interesting structure. However, in tracking changes over time, we need to be able to find corresponding communities in clusterings taken from the data at different points in time. If the clusterings are very sensitive to small perturbations of the input data, distinguishing between "real" changes versus "accidental" changes in the higher-level structure becomes difficult, if not impossible. In the clusterings of our linked network data, we found there are a large number of relatively random clusters that do not correspond to real community structures. These random clusters obscure the real temporal changes. Fortunately, we found that, when performing a series of agglomerative clustering runs, each run on slightly perturbed input data, one can identify a stable set of clusters that occur in a significant proportion of the clusterings. Moreover, these stable clusters appear to correspond to the true underlying community structure of the network. We refer to such stable clusters as natural communities. We use the notion of natural communities to show that we can track these natural communities effectively over time, and can therefore characterize the temporal evolution of the network.

Data Set

We used an October 2001 snapshot of the NEC CiteSeer database (5). At that time, the CiteSeer database contained the full text and bibliographies of ≈250,000 papers. These are mostly related to computer science, with a small collection covering other topics like physics, mathematics, and economics. The papers are mostly published after 1990, and the set is growing by ≈25,000 papers per year. In addition, the database contains title and author information on another 1.6 million earlier papers that are referenced by the 250,000 set but whose full text is not contained in the database.

We analyze the citation graph induced by this data set: vertices correspond to all 1.85 million papers in the database; there is a directed edge from paper A to paper B if A references B. We call the set of 250,000 papers whose full-text and bibliography are known the core of the citation graph. The papers in the core have citations to each other and to the 1.6 million earlier papers. We do not have the reference lists for the papers outside the core. So, their out-degree is 0, whereas their in-degree is at least 1. Fig. 1 gives a pictorial representation of our graph, and Table 1 contains key statistics. The out-degree (number of papers in the bibliography of a paper) of a typical node ranges from 5 to 25. The median out-degree for the core papers is 14. Interestingly, the majority of core papers are uncited (in-degree = 0). Refs. 6 and 7 describe methods for removing inaccuracies in the CiteSeer citation graph caused by the automatic generation of the graph.

The basic statistics of this graph already reveal that its structure is very different from a standard random graph. About 1 in every 100 papers receives >20 citations, 1 in every 1,000 papers has 300 citations or more, and 18 papers of the 1.85 million have >1,000 citations. This pattern is indicative of the heavy-tailed nature of the data, characterized by a power law in the in-degree (8). An interesting research question concerns the role of the highly cited papers. For example, are such nodes essential in the definition of the hidden community structure or does such structure remain even after removing high degree nodes from the graph? Also, are such nodes essential in the formation of new communities?

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

§To whom correspondence should be addressed. E-mail: selman@cs.cornell.edu.

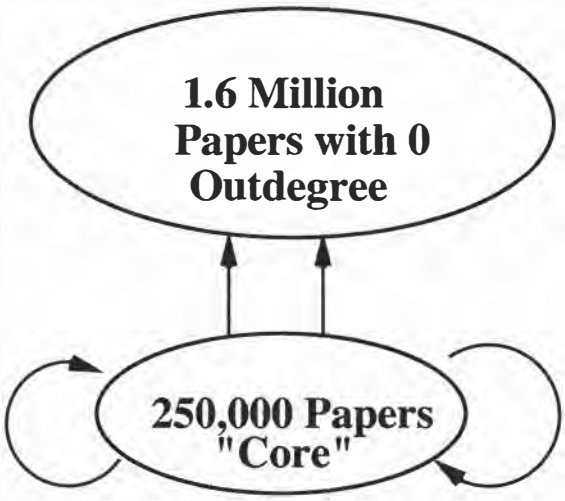


Fig. 1. Structure of NEC CiteSeer citation graph.

Instabilities and Natural Communities

Hierarchical agglomerative clustering starts with each paper in a cluster by itself. At each stage, the two "closest" clusters are merged. The process is repeated until all papers are in a single cluster. The overall process results in a "clustering tree" (referred to as a dendrogram in much of the literature), with the single paper clusters at the leaves. Each internal node corresponds to a cluster resulting from merging its two children.

Researchers have used many different distance measures. We believe that the natural community concept is valid independent of the distance measure used and thus we selected one based on cosine similarity, which is the standard similarity measure in the literature (9). With each paper p, we associate an N-dimensional reference vector r_p, where N is the total number of papers in the CiteSeer database ($N = 1.85$ million). There is a one in element i of r_p if p references paper i, otherwise the entry is 0. The similarity between two papers p and q can now be measured in terms of the cosine of the angle between the associated reference vectors, r_p and r_q. More formally, the similarity of p and q is defined to be

$$similarity(p, q) = \cos(r_p, r_q) = \frac{r_p \cdot r_q}{\|r_p\| \, \|r_q\|}, \qquad [1]$$

where $r_p \cdot r_q$ represents the inner product of r_p and r_q and $\|r_p\|$ represents the length of vector r_p. So, if two papers have no references in common, then their similarity is minimal, i.e., 0 (90° angle); two papers citing exactly the same set of papers have maximal similarity, i.e., 1 (0° angle). To get a distance measure between papers, we simply use 1− the cosine, so the distance between papers ranges from 0 to 1. When merging two papers or clusters, we represent the new cluster by the normalized sum of all of the individual papers' reference vectors, called the "centroid" of the cluster. (Our clustering method is thus a standard centroid-based agglomerative clustering technique based on cosine similarity; ref. 10.) For a cluster containing a single paper, the centroid is simply the reference vector of the paper itself. Finally, we define the distance between two clusters C and C'. Let n_C and $n_{C'}$ be the number of papers in each cluster, and let r_C and $r_{C'}$ be the centroids of the clusters. Then

$$distance(C, C') = \sqrt{\frac{n_C n_{C'}}{n_C + n_{C'}}} (1 - \cos(r_C, r_{C'})) \qquad [2]$$

The square root scaling factor is used to force smaller communities to merge together before larger ones (11). This particular scaling factor leads to well balanced merge trees.

Our distance measure is a form of bibliographic coupling (12). A prominent alternative is to use cocitation analysis (13). In cocitation, two papers are judged similar if they are both cited by another paper. This is a very useful similarity measure. However, for this measure to work properly, a certain time-lag is required in order for papers to build up a citation record. Because our objective is to detect changes as early as possible, we opted for the common reference set approach. This also allows us to group papers that are not cited at all or only rarely cited, which is a significant portion of all papers.

To verify that the clustering algorithm and distance function were satisfactory, we compared the quality of the clusters we obtained to clusters obtained by standard techniques such as k means. One method of comparison is to count the number of journals and conferences needed to cover 90% of the papers in a cluster. The assumption here is that most journals, with a few exceptions, such as *SIAM Review*, are on a focused topic. Thus, the fewer the number of journals needed to cover a cluster, the tighter the cluster. In our tests, we found that the clusters obtained with the agglomerative algorithm were better defined than the clusters obtained by other methods.

Instabilities. To determine the set of natural communities, we examine changes in the agglomerative clustering trees under minor perturbations of the input data (14). More specifically, we compare different clusterings of the CiteSeer data, where we remove a small randomly selected set of papers (5%) before each clustering run. Given a base tree T_1, we compare how well the clusters in T_1 match with those in a second tree T_2, obtained on a different clustering run.

Let C and C' be two clusters of papers we wish to compare. Treating C and C' as sets, we define a value match(C, C') (between 0 and 1), as follows:

$$\text{match}(C, C') = \min\left(\frac{|C \cap C'|}{|C|}, \frac{|C \cap C'|}{|C'|}\right). \qquad [3]$$

The definition ensures that a high match values (close to 1) occurs when two clusters have many papers in common and are roughly of the same size. We define the value best-match (C, T) as the highest match (C, C') value for any cluster C' in T.

We considered a total of 45 clusterings of the CiteSeer graph. Each run uses a graph with a random 5% of the core papers removed. Full clustering runs on a data set of this size require

Table 1. Statistics of CiteSeer citation graph

Data set	n
Nodes	1,859,659
Nodes core	252,493
Edges	4,584,756
Average out-degree core	18
Median out-degree core	14
Median in-degree core	0

Table 2. Best-match values

Size range	No. of clusters in base tree	Average best-match	SD
100–400	2,812	0.42	0.07
401–1600	558	0.41	0.07
1,601–6,400	149	0.38	0.07
6,401–102,400	46	0.40	0.08

Table 3. Natural communities

Size range	No. of natural communities	Average best match	SD
100–400	116	0.74	0.05
401–1600	32	0.62	0.07
1,601–6,400	17	0.60	0.06
6,401–102,400	5	0.60	0.10

an efficient algorithm, and so we carefully exploit the sparseness of the underlying network. This allows for an efficient update of the set of intercluster distances after each merge. Code and data are available on request.

Table 2 gives the average best-match values of the clusters in the base tree T_1 matched against the other 44 trees. So, for example, the first row in the table shows that T_1 has 2,812 clusters containing between 100 and 400 papers. For each cluster C in this size range, we found the best matching cluster and its best-match value in trees T_2 through T_{45}. The average over these best-match values is 0.42 with a standard deviation of 0.07.

Table 2 shows that the average cluster matches quite poorly to its closest match in the other tree (average best-match value only $\approx$0.40). Interestingly, we can take advantage of these instabilities, because these clusters are not uniformly unstable and therefore can be exploited to uncover the true hidden structure of the data. In fact, a careful examination of the results of many runs shows that a small number of clusters, $\approx$170, appear in a good fraction of clusterings, and it is these clusters that correspond to recognizable topics. These "fixed points" in the CiteSeer graph are what we call natural communities, and these are the communities whose evolution we will track over time.

Natural Communities. We define natural communities as follows. We fix an input data perturbation value of 5%. Then we produce a series of subgraphs $G_1, G_2, \ldots, G_n$ of the original network G (the CiteSeer citation graph) where each G_i is the subgraph of G induced by a random subset of 95% of the core vertices of G. Our clustering algorithm then produces a set of clustering trees $\mathbf{T} = \{T_1, T_2, \ldots, T_n\}$. We choose the first tree T_1 as our base tree. We now define a natural community or cluster as follows.

Definition 3.1. A community C in base tree T_1 is natural *iff* in a fraction f of the clustering trees in $\mathbf{T}$ the best-match of C has a value greater than p, a predefined threshold.

The definition has two parameters: f, the fraction of trees out of n trees total, and p, a lower-bound for the best match. Depending on what values one chooses for these parameters, one obtains more or less well defined natural communities. In practice, we set these values sufficiently high to select clusters that are clearly different from the average cluster in the tree.

Using $n = 45, f = 0.6, p = 0.7$ for clusters with <1,000 papers and $p = 0.5$ for larger clusters, we found 170 natural communities of size 100 or greater in the CiteSeer graph, covering all aspects of computer science and portions of other fields like math and physics (see Table 3). These natural communities were selected from >3,500 clusters with size >100 in the base tree. Note that these natural communities vary in strength and their precise number depends, of course, on the setting of f and p. By using keyword data and journal titles, we found the natural communities to be quite coherent. In particular, the smaller to medium natural communities (up to a few thousand papers) correspond to well defined areas. Some example communities are listed in Table 4 (more details below). [Smaller natural communities are better defined. By using a different, but somewhat nonstandard, distance measure, one can also obtain better defined larger natural communities; ref. 14.]

We now turn to our main objective: the use of these natural communities in tracking the temporal evolution of the network.

Tracking Natural Communities

The key question remaining is how well the natural communities allow us to track the temporal evolution of the community structure in our network data.

In particular, we need to validate that when the network evolves over time and a few years of papers are added (*i*) there is not a dramatic shift in terms of natural communities, and (*ii*) that the occurring changes have a plausible interpretation in terms of the evolution of the field. These are inherently empirical questions. The results discussed below will show that our notion of natural communities satisfies both criteria, thereby making the concept a good candidate for use in temporal tracking in large networked data sets.

Method. To study the temporal evolution process in detail, we will track changes for a subset of the natural community data described above. We use two snapshots: the time periods 1990–1998 (referred to as the 1998 data set) and 1990–2001 (referred to as the 2001 data set). As such, our goal is to study changes in community structure as they occurred during the three years from 1999 to 2001. [The core set of the CiteSeer data set consists of papers available in digital form on the web. The literature coverage of the earlier years of the collection is less complete because the fraction of papers available on the web was limited, but by the late 1990s, the coverage for computer science had become quite comprehensive.]

In the 2001 data set, there are $\approx$100 natural communities containing between 100 and 350 papers. To analyze temporal changes in detail, we considered a subset of 20% of these communities (18 total) for closer analysis. Our selection of communities was representative of the overall set of communities in terms of size and year distribution. The communities contained 3,200 papers total. Let P_{2001} be the set of these papers. We create a citation subgraph containing only papers from P_{2001} and the references in these papers. We also removed some

Table 4. Established natural communities

Topic	Size in 1998	Size in 2001	Percentage in 2001
Digital watermarking	97	172	35.5
Data mining and association rules	78	128	25.0
Game search trees and artificial intelligence	161	172	8.7
Network traffic control	237	258	8.5
Crash recovery for distributed systems	139	151	7.3
Asynchronous circuit designand verification	231	244	6.6
Synchronous and asynchronous systems	203	219	6.4
Complexity theory: enumerability and querying	78	84	6.0
Query optimization for parallel databases	119	125	4.0
Fractal image coding and compression	86	89	2.2

Table 5. Emerging natural communities

Community	Size	Percentage in 2001
1998		
Networking (two communities)	237 + 130	
Quantum complexity	96	
2001		
Ad hoc/wireless networks	130	49.2
Quantum computation	140	30.0
Subcommunities		
Quantum complexity	82	15.9
Quantum algorithms and communication	38	76.3

low-quality information: all core papers that reference fewer than five other papers and all noncore papers only referenced once. This reduced the size of the subgraph by ≈20%. We repeat this procedure to create our 1998 graph by starting with papers up to and including 1998 from the set P_{2001} (a subset of 2,791 papers).

Results. We determined the natural communities for each data set by considering 10 clusterings for each graph and by using $f = p = 0.8$ in our definition of natural communities. We considered all natural communities with at least 75 papers. We then compared the natural community trees for the 1998 and the 2001 data, by finding for each 1998 natural community the best matching natural community in our 2001 data and vice versa.

Our first observation is that most of the 1998 communities have a good match (at least 70%) with a 2001 community (and vice versa). (Note that the 2001 data set contains ≈13% more papers than the 1998 set, with some communities growing by >30%.) Also, the natural community tree structures largely match up. Based on the matching data and the trees, we classify the natural communities in the 2001 data set as either established or emerging.

The established communities are given in Table 4. The table gives the size of the communities in 1998 and 2001 and then the percentage of papers in the 2001 community that appeared after 1998 (indicating the growth rate). The topic of each community was determined by considering the most frequent content words in the titles of the papers in each community. The communities are sorted by growth rate. We see that the growth rates vary quite a bit: some communities are very active and growing fast, such as digital watermarking and data mining, but several other communities appear stagnant, such as fractal image coding and compression, and query optimization for parallel databases.

From the perspective of temporal evolution, the most interesting changes involve the emergence of new communities. We identified two emerging communities: ad hoc/wireless networks and quantum computing. (In ad hoc networks, one studies self-configuring, distributed networks, generally wireless.) See Table 5. These emerging communities are consistent with recent developments in the field.

Wireless networks. Our first example is the emergence of the wireless community. In 1998, we have two natural communities centered around "network systems" with 367 papers. These communities consist of a combination of optical networking, distributed computing, and crash recovery papers with some initial papers on ad hoc/wireless networks. However, at this time, there is no well defined community on ad hoc/wireless networks. However, there is a significant change in networking papers over the 1999–2001 period, as ≈60 papers on ad hoc/wireless networks are added to the database. As a result, we find that, in the 2001 data, the ad hoc/wireless papers form a distinct natural community consisting of ad hoc/wireless papers from the 1998 set with the post-1998 papers added. In the 2001 cluster tree, this new community merges in with the larger network community at a higher level.

Quantum computing. A second example is the emergence of the quantum algorithms and communication community within quantum computing. This is an example of a community that is branching out over time (i.e., it is an evolving community). In the 1998 set of natural communities, we find that there is a natural community of size 96 that contains papers on quantum computing and complexity theory. In the 2001 set, this community has grown to 140 papers. However, the 2001 clustering now reveals further substructure: there are two distinct subcommunities of the size 140 community: one on quantum complexity (size 82; fairly stable) and another, fast growing community of 38 papers (20 more papers merge in separately). After examining the titles, it is clear that most of these papers cover quantum algorithms and quantum communication, both very hot topics in the past few years. So, in 1998, the quantum community was mostly centered on one topic; in 2001, the community was branching and growing quickly (theory conference agendas actually reflect this). Given the recent explosion of work in the area of quantum computing, it is encouraging to see these developments reflected in our natural community data.

In summary, these examples show that our notion of natural communities provides a promising tool for studying the temporal evolution of linked networks.

Related Work

Early pioneering work on discovering scientific communities using reference linkage information was done by Small and colleagues (13, 15). More recently, the NEC CiteSeer group succeeded in identifying intellectual communities in the CiteSeer database by using new variants of cocitation analysis (16) and network flow methods (17). The main impetus for the recent renewed activity in this area comes from the increasing importance of large linked networks in general, not just networks based on citation data (e.g., ref. 18). Indeed, recent work (19) explores the dynamics of social networks by simultaneously analyzing coauthorship and citation networks. For future work, it would be interesting to consider the relationship between authorship and natural clusters of papers as we identified here.

A key aspect that distinguishes our work is the emphasis on the temporal evolution of the network. As a consequence, for example, cocitation is less useful as a similarity measure, because it takes time to build up a cocitation record. Similarly, the use of highly cited papers, as in ref. 16, to identify core communities, also has limitations when looking for the most recent changes in the network involving emerging communities, because again it takes time to build up a citation record. (Ref. 16 measures the activity level of established communities by considering growth rates.) A related question is whether so-called hubs and authorities, as introduced by Kleinberg (20), form quickly enough to track recent changes. In general, a more detailed comparison between our natural communities and communities identified by using these other approaches is needed.

Another aspect that differentiates our work is its focus on stability: to track clusters over time, it is important that the clustering hierarchy be relatively stable. In many other applications, what matters most is not stability but finding an organization of items that, on human inspection, is coherent. With respect to the CiteSeer document collection, this amounts to identifying key computer science topics, such as systems and databases using titles and abstracts (17). However, the importance of stability is gaining recognition. Refs. 21 and 22 analyzed the stability of the two popular link-based ranking algorithms, HITS (20) and PageRank (23). They point out that intuitively we would not want the rankings for a given query to change much if the base data set, for example, the World Wide Web, is altered

slightly. They go on to develop algorithms that stabilize the HITS rankings.

Conclusions

We have provided a framework for studying the temporal evolution of the community structure of large linked networks. The notion of natural communities can be used to identify a relatively stable core of a hierarchical agglomerative clustering. Our approach exploits the inherent instabilities in clusterings in high-dimensional spaces (24). The true structure in the data are revealed by averaging out the large number of "accidental" clusters that emerge in any single clustering run. In our experiments on the CiteSeer network, we showed how the natural communities can be used to study the evolution of the network by tracking established communities and uncovering new, emerging community structure. Our next step is to evaluate our approach on other evolving linked networks.

We thank Steve Lawrence for making the October 2001 snapshot of the NEC CiteSeer database available to us. We thank Richard Shiffrin and Katy Börner for valuable feedback on an earlier version of this paper. We also thank Justin Yang for assistance with the experiments. This work was supported in part by National Science Foundation CAREER Award IIS-9734128, an Alfred P. Sloan Research Fellowship, National Science Foundation Information Technology Research Grant IIS-0312910, and the Intelligent Information Systems Institute at Cornell University sponsored by Air Force Office of Scientific Research Grant F49620-01-1-0076.

1. Watts, D. J. & Strogatz, S. H. (1998) *Nature* **393,** 440–442.
2. Watts, D. (2003) *Six Degrees: The Science of a Connected Age* (Norton, New York).
3. Barabàsi, A.-L. (2002) *Linked: The New Science of Networks* (Perseus, New York).
4. Erdös, P. & Rényi, A. (1960) *Pupl. Math. Inst. Hungarian Acad. Sci.* **7,** 17–61.
5. Giles, C. L., Bollacker, K. D. & Lawrence, S. (1998) in *Proceedings of the International Conference on Digital Libraries*, eds. Witten, I., Akscyn, R. & Shipman, F. M. (Assoc. Comput. Machinery Press, New York), Vol. 3, pp. 89–98.
6. Cohen, W., Kautz, H. & McAllester, D. (2000) in *Proceedings of the Association of Computational Machinery Special Interest Groups Knowledge Discovery in Data and Data Mining*, eds. Ramakrishnan, R., Stolfo, S., Bayardo, R. & Parsa, I. (Assoc. Comput. Machinery Press, New York), Vol. 6, pp. 255–259.
7. Pasula, H., Marthi, B., Milch, B., Russell, S. & Shpitser, I. (2003) in *Advances in Neural Information Processing Systems*, eds. Becker, S., Thrun, S. & Obermayer, K. (MIT Press, Cambridge, MA), Vol. 15, pp. 1401–1408.
8. Adler, R. J., Feldman, R. E. & Taqqu, M., eds. (1998) *A Practical Guide to Heavy Tails* (Birkhäuser, Boston).
9. Salton, G. (1989) *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer* (Addison–Wesley, Boston).
10. Jain, A. K. & Dubes, R. C. (1998) *Algorithms for Clustering Data* (Prentice-Hall, Upper Saddle River, NJ).
11. Duda, R. O. & Hart, P. E. (1973) *Pattern Classification and Scene Analysis* (Wiley, New York).
12. Kessler, M. M. (1963) *Am. Document* **14,** 10–25.
13. Small, H. (1973) *J. Am. Soc. Info. Sci.* **24,** 265–269.
14. Hopcroft, J., Khan, O., Kulis, B. & Selman, B. (2003) in *Proceedings of the Association of Computational Machinery Special Interest Groups Knowledge Discovery in Data and Data Mining*, eds. Ramakrishnan, R., Stolfo, S., Bayardo, R. & Parsa, I. (Assoc. Comput. Machinery Press, New York), Vol. 9, pp. 541–546.
15. Small, H. & Griffith, B. C. (1974) *Sci. Stud.* **4,** 17–40.
16. Popescul, A., Flake, G., Lawrence, S., Ungar, L. & Giles, C. L. (2000) *Advances in Digital Libraries, ADL 2000* (IEEE, New York), pp. 173–182.
17. Flake, G. W., Lawrence, S. & Giles, C. L. (2000) in *Proceedings of the Association of Computational Machinery Special Interest Groups Knowledge Discovery in Data and Data Mining*, eds. Ramakrishnan, R., Stolfo, S., Bayardo, R. & Parsa, I. (Assoc. Comput. Machinery Press, New York), Vol. 6, pp. 255–259.
18. Gibson, D., Kleinberg, J. M. & Raghavan, P. (1998) *Proc. Hypertext 1998 Conf.* **9,** 225–234.
19. Börner, K., Maru, J. T. & Goldstone, R. L. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5266–5273.
20. Kleinberg, J. M. (1999) *J. Assoc. Comput. Machinery* **46,** 604–632.
21. Ng, A. Y., Zheng, A. X. & Jordan, M. (2001) *Proc. Int. Joint Conf. Artificial Intelligence* **17,** 903–910.
22. Ng, A. Y., Zheng, A. X. & Jordan, M. (2001) *Proc. Assoc. Comput. Machinery Spec. Interest Groups Inf. Retrieval Conf., New York* **24,** 258–266.
23. Page, L. & Brin, S. (1998) *Comput. Networks ISDN Syst.* **30,** 107–113.
24. Aggarwal, C. C., Hinneburg, A. & Keim, D. A. (2001) in *Lecture Notes in Computer Science*, eds. Van den Bussche, J. & Vianu, V. (Springer, Heidelberg), pp. 420–434.

Traffic-based feedback on the web

Jonathan Aizen*†, Daniel Huttenlocher*, Jon Kleinberg*‡, and Antal Novak*

*Department of Computer Science, Cornell University, 4130 Upson Hall, Ithaca, NY 14850; and †Internet Archive, Presidio, San Francisco, CA 94129

Usage data at a high-traffic web site can expose information about external events and surges in popularity that may not be accessible solely from analyses of content and link structure. We consider sites that are organized around a set of items available for purchase or download, consider, for example, an e-commerce site or collection of online research papers, and we study a simple indicator of collective user interest in an item, the batting average, defined as the fraction of visits to an item's description that result in an acquisition of that item. We develop a stochastic model for identifying points in time at which an item's batting average experiences significant change. In experiments with usage data from the Internet Archive, we find that such changes often occur in an abrupt, discrete fashion, and that these changes can be closely aligned with events such as the highlighting of an item on the site or the appearance of a link from an active external referrer. In this way, analyzing the dynamics of item popularity at an active web site can help characterize the impact of a range of events taking place both on and off the site.

Large information repositories are often studied not just in terms of their content, but also in terms of the structures that grow up around this content. In the scientific literature, the network of citations provides a clear example of this type of structure; it can supplement the text of published papers by highlighting work that others have found to be important. This principle extends naturally to the information contained in web sites as well; hyperlinks on the web provide a powerful framework for organization and analysis that parallels the use of citations and cross-references in other media (for an example, see ref. 1).

Web sites and other online documents, however, can be further annotated with information typically not available in traditional print sources: the patterns of usage generated by visitors to the site. At the most basic level, explicit analysis of a web site's usage can play a role similar to that of hyperlink analysis; for instance, uncovering parts of the site that have attracted large numbers of visitors can help to highlight important content for future users. But usage data are considerably more dynamic and volatile than link structure; usage changes quickly in response to external events and surges of popularity, many of which are significant but too transient to leave behind a long-term mark on the site. With effective means for analyzing this usage dynamics, we can thus characterize a web site along a dimension that neither content nor link structure is able to capture.

Our work is based on an analysis of usage data from the Internet Archive (www.archive.org), which maintains a large collection of downloadable media, including movies, music, and books, as well as snapshots of the web itself reaching back to its early history. Our approach, however, is applicable to a wide range of web sites offering items that users may or may not want to acquire (e.g., for sale or download). Such sites typically contain three distinct types of content: navigational structure, item descriptors (an individual page associated with each item, providing a description of the item together with the option to acquire it), and the items themselves. This kind of navigation–description–acquisition structure is common in e-commerce sites, such as amazon.com, and in online libraries or research paper collections such as the e-print arXiv (2) and CiteSeer (3). In the case of the Internet Archive, this structure is manifested through a "details" page for each media item, containing a summary of the content together with user reviews and links for downloading the item.

In the following sections, we develop methods for modeling and tracking the popularity of items at web sites with this structure. We introduce the batting average, the proportion of visits that lead to acquisitions, as a measure of an item's popularity, and we illustrate why it is a useful complement to traditional measures such as visit or acquisition counts alone. We then develop a stochastic model of how the batting average varies over time, and we use it to examine the level of interest in certain items in the Internet Archive. We find that many of the changes in item popularity have a discrete nature, they occur suddenly, and their onset can be related to specific events taking place either on or off the site. Further, we argue that knowledge of these changes and the events surrounding them can be of value both to users of the site and the site's administrators.

The Batting Average of an Online Item

There are several quantitative ways to try to capture an item's popularity. Consider, first, ranking each item in order of its acquisition count, the number of times it has been acquired (e.g., downloaded or purchased). Many web sites offer this type of ranking to users in the form of a "most popular" list. Such lists, while clearly providing useful feedback, suffer from two intrinsic (and related) problems: they typically change very little over time, because the top items on these lists build up large counts that are relatively impervious to localized trends, and they are self-reinforcing in the sense that users are often driven to look at an item simply because it appears on one of these lists.

We have been studying an alternative measure, the batting average; although still simple to define, it exhibits a more complex dynamics. On any web site with a description–acquisition structure, the batting average of an item is defined as the number of acquisitions of the item divided by the number of visits to its description. Thus, the batting average can be thought of as a kind of inherent "appeal" of an item, the probability that a visit will lead to acquisition, averaged over all visitors to the item's description.

Both the acquisition count and the batting average have the potential to change significantly when an item is highlighted in some way, either on the site or by an active off-site referrer, and is thereby exposed to a larger or different population of users. The way in which these two measures generally experience change, however, is quite different. The acquisition count never

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

Abbreviation: HMM, hidden Markov model.

‡To whom correspondence should be addressed. E-mail: kleinber@cs.cornell.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0307539100

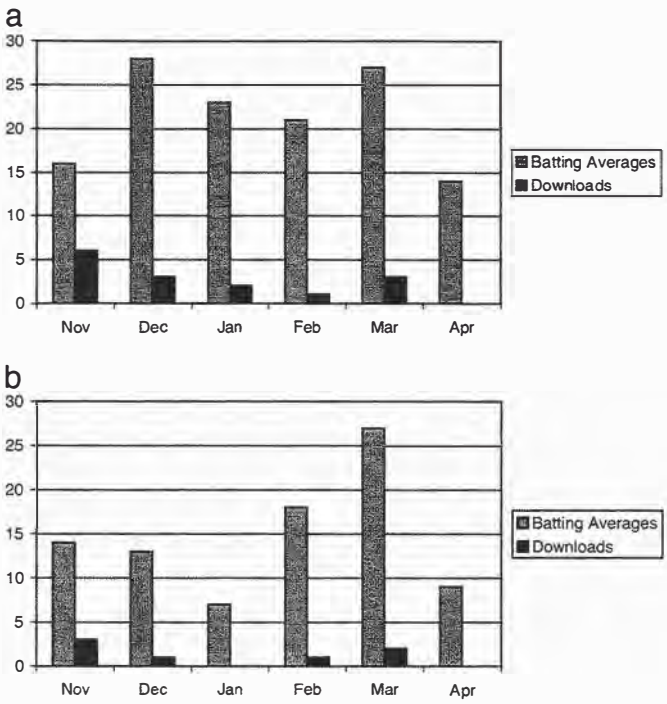


Fig. 1. The number of changes per month in top audio (a) and movie (*b*) items: ranking by batting average versus ranking by downloads.

decreases with increasing exposure of the item, and the magnitude of the increase in large part reflects the extent of the exposure. The batting average, on the other hand, may go up or down when the item is exposed to a new population, depending on whether this new population contains a larger or smaller fraction of users who are interested in acquiring it; in this way, the change to the batting average reflects something about the interests of the new population relative the item's standard set of visitors. More generally, the dynamics of an item's batting average over time can help one to dissect the mix of users who encounter and evaluate it at different times and for different reasons.

As one concrete indication of the different dynamics exhibited by the acquisition count and the batting average, we consider the effects of reporting these quantities as feedback to users. Because its media collections became public, the Internet Archive site has featured continuously updated lists of the items with the highest acquisition count, displayed separately for movies, audio, and texts. Beginning in November 2002, on the same pages, lists of the items with the highest batting averages (corrected for small sample sizes) were added. We find that the lists of most acquired items change very infrequently, reflecting their self-reinforcing or "rich-get-richer" character: users are driven to look at (and then often acquire) these items simply because of their presence on the lists, which increases their acquisition count. The lists of items with the highest batting averages, on the other hand, are significantly more "turbulent": when an item enters such a list, its visibility on the site increases significantly, and a broader population of users is driven to look at its description. At this point the item's batting average can remain stable or increase if it is of mainstream interest or else the batting average will drop and the item will rapidly leave the top-ranked list. Fig. 1 shows the number of days in each month (November 2002 through April 2003) when the "top-5 acquired" and "top-5 batting averages" lists experienced a change. The greater turbulence of the batting average list is reflected in the fact that it changes every 2–3 days on average, as opposed to every couple of weeks.

We have also observed that the distribution of acquisition counts across all items at the Internet Archive has a heavy-tailed distribution, whereas the distribution of batting averages does not. Heavy-tailed distributions are widely observed in settings that are dominated by rich-get-richer dynamics (4); this is a further quantitative reflection of the contrast between acquisition count and batting average.

Tracking Interest over Time

We have argued that aggregate interest in an item, as measured by the batting average, may change whenever the item is exposed to a new mix of users. Moreover, if we think about the potential causes of an item's increased exposure, many of these are not gradual trends but discrete events, occurring at precise points in time. Consider, for example, the effect of highlighting an item on a top-level page on the site, or the effect of a new link from an active off-site referrer; this highlight or hyperlink first appears at a specific moment, and the item's batting average is particularly susceptible to change at such moments. With a means for identifying discrete changes in the batting average, we can assess the extent of this phenomenon in practice and automatically identify the most significant events that affect interest in each item in the collection.

Thus, we need a way to meaningfully express the "instantaneous batting average" of an item at any point in time, so that we can identify the moments when this quantity changes. Defining such an instantaneous measure is a bit subtle, because at any one particular point in time, we simply have information about a single user's decision to download the item or not. One simple approach would be to average the results of a number of consecutive user visits, obtaining a batting average over a "sliding window" in time, but as we discuss further below, we have found that this is not effective at localizing a small set of changes caused by external events. Instead, we make use of the stochastic modeling framework of hidden Markov models (HMMs) (5), explicitly representing the underlying download probability as a "hidden state" in the process, and identifying the moments when this state changes.

To motivate this, consider first a simple model of an item's batting average: a sequence of users visit the item's description, and each user independently decides whether to acquire the item by flipping a coin of bias b (which, over a long enough sequence of users, will be approximately equal to the observed batting average.) We now consider a richer model in which the underlying bias of the coin can change. Thus, there is an underlying set of possible coin biases $0 < b_1 < b_2 < \cdots < b_n < 1$, which we view as the potential states of the process. Users arrive at discrete time steps $t = 1, 2, \ldots, T$ and at time t, the decision to download is made with a bias of b_{i_t} (where $0 < i_t < n + 1$). After each step, there is some probability that the bias will change; specifically, there is a function $\gamma(\cdot, \cdot)$ so that, if the current bias at visit t is b, it will change to b' at visit $t + 1$ with probability $\gamma(b, b')$.

The decision by each of the T visitors to download the item or not can be encoded as a length-T sequence $\mathbf{d} = (d_1, d_2, \ldots, d_T)$ of 0s and 1s. Our goal is to find the corresponding sequence of biases $\mathbf{b} = (b_{i_1}, b_{i_2}, \ldots, b_{i_T})$ that is most likely given download sequence $\mathbf{d}$; in other words, we want to maximize $\Pr[\mathbf{b}|\mathbf{d}]$. By Bayes' theorem, this is equivalent to maximizing $\Pr[\mathbf{d}|\mathbf{b}]\cdot\Pr$. The first term decomposes into a sequence of independent download decisions, $\Pr[\mathbf{d}|\mathbf{b}] = \Pi_{t=1}^{T} \Pr[d_t|b_{i_t}]$, where each factor is simply the probability of a 0 or 1 given the bias: $\Pr[1|b_{i_t}] = b_{i_t}$ and $\Pr[0|b_{i_t}] = 1 - b_{i_t}$. The second term factors into a sequence of probabilistic transitions according to our model, $\Pr[\mathbf{b}] = \Pi_{t=1}^{T-1} \gamma(b_{i_t}, b_{i_{t+1}})$. Finally, it is useful to take the negative logarithm of the expression, so that we are seeking to minimize a sum rather than maximize a product:

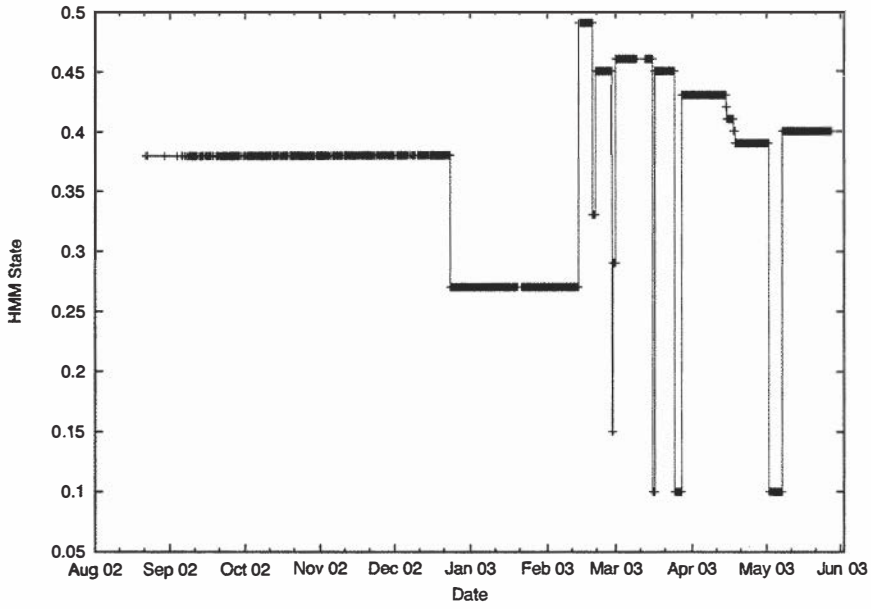


Fig. 2. Tracking the batting average of *What You Should Know About Biological Warfare* as a function of time, using state transitions in a HMM.

$$-\log(\Pr[\mathbf{d}|\mathbf{b}]\cdot\Pr[\mathbf{b}])$$

$$= \sum_{t=1}^{T} \left\{ \begin{array}{ll} -\log b_{i_t} & \text{if } d_t = 1 \\ -\log(1 - b_{i_t}) & \text{if } d_t = 0 \end{array} \right\} + \sum_{t=1}^{T-1} -\log \gamma(b_{i_t}, b_{i_{t+1}}). \quad [1]$$

In our case, we use a set of biases $\{b_i\}$ that range from 0.10 to 0.90 in increments of 0.01. We limit the range of biases to avoid values near 0 and 1, where the first term of the summation (corresponding to $\Pr[1|b_{it}] = b_{it}$ and $\Pr[0|b_{it}] = 1 - b_{it}$) becomes unbounded. We choose a step size of 0.01 because empirically a smaller step size of 0.001 yields essentially the same results, whereas a larger step size of 0.1 fails to distinguish a number of events.

The quantity $-\log \gamma(b, b')$ in the second term of Eq. **1** can be viewed as a state transition cost (because we seek to minimize it), and we model it as increasing linearly in relation to the distance between the biases b and b':

$$-\log \gamma(b, b') = \min(k_1|b - b'|, k_2|b - b'| + k_3),$$

where the k_i are positive constants with $k_1 > k_2$; these constants are chosen to yield a small number of state transitions per item ($\approx$10–20) in practice. The rationale for the two-slope model underlying the state transition cost is that we want to be able to detect sudden large changes in bias that result from discrete external events such as adding a new link, while at the same time not having the model drift from one bias to another and back again. Thus the higher slope for small changes encourages the system to stay in a given state rather than bouncing around, but at the same time large changes in bias do not incur much more of a penalty than moderate changes. This is a standard kind of "truncated model" used in robust statistics. Many other cost functions, including quadratic (which would correspond to Gaussian distributed transition probabilities) could also be used here.

Intuitively, the two terms in Eq. **1** reflect opposing forces in our tracking of the batting average over time: the first term seeks to produce a sequence of biases that accurately follows each individual download decision, whereas the second term seeks to produce a sequence of biases that is relatively constant, reflecting the notion that discrete changes happen only rarely. Minimizing this expression thus corresponds to a balance between these two qualitative goals. The standard Viterbi algorithm (5) can find the state sequence minimizing Eq. **1** in time $O(Tn^2)$. Because we are dealing with extremely long sequences and using a relatively large set of biases, this running time would be prohibitive; to avoid this problem, we use a recently developed algorithm that exploits the structure of the state transition costs to perform the minimization in time $O(Tn)$ (6, 7).

The result of this computation is a sequence of biases that reflects our best estimate of the instantaneous batting average across the full sequence of user visits. With a setting of parameters under which this quantity changes infrequently, these discrete moments of change become natural breakpoints around which to understand the evolution of interest in the item. Because of the large set of possible biases considered by the algorithm, we find that the changes tend to correspond closely to recognizable events involving the site; in contrast, models with a sparser set of possible biases (i.e., with a smaller value of n) produce state transitions that do not track these events well. We have also considered the use of stochastic models for analyzing the rate of item downloads and the rate of visits to item description pages. This approach is analogous to our stochastic analysis of batting averages, but with several differences. First, because download and visit rates are measured in users per unit time, it is natural to use a probabilistic waiting-time model; such a model is more difficult to meaningfully discretize into states than the biases in our coin-flipping model for batting average, which naturally reside in the interval [0,1]. Second, as discussed in the previous section, changes (captured as state transitions) in the batting average tell us about the varying interest level of the underlying population in a way that state transitions corresponding to spikes in the download or visit rate alone do not. For these reasons, we leave the stochastic analysis of download and visit rates outside the scope of the present article.

Before we analyze the set of batting average state transitions systematically, it is useful to consider a single item in some detail, to get a qualitative sense for what one finds from temporal changes in the batting average. Thus, in Fig. 2 we plot the HMM state for the batting average (the value of the hidden bias

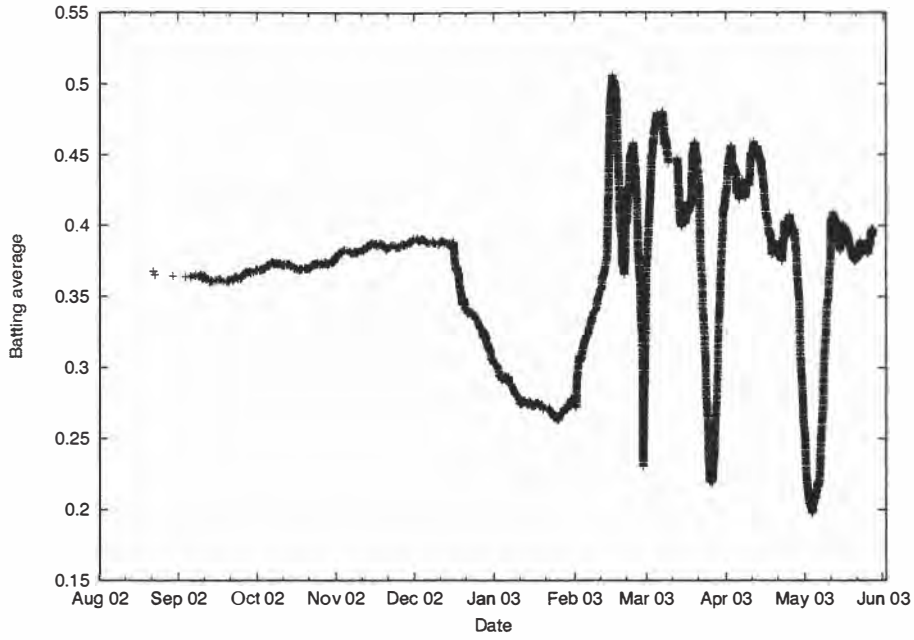


Fig. 3. A noisier method of tracking the batting average for *What You Should Know About Biological Warfare*, based on a sliding window (Gaussian convolution) of contiguous visits.

variable) as a function of time for the Internet Archive's online copy of the 1952 civil defense film *What You Should Know About Biological Warfare*. Roughly, we see that the batting average begins at a high level (≈0.38), then drops to a lower level (≈0.26), and returns to a higher level again (between 0.40 and 0.50), with the final higher period interrupted by five brief, sharp drops.

Annotating these transitions in terms of events both on and off the site, a clear picture of the item's history emerges. The initial drop to a lower level in December 2002 occurred when the item was added to the Pick List, an unannotated list of (recommended) titles on a top-level page at the Internet Archive. The subsequent return to a higher level in February 2003 occurred when the item was moved (a week after Colin Powell's testimony on biological weapons to the United Nations Security Council) from the Pick List to the Collection Spotlight, a more extensive means of highlighting in which the title is accompanied by a brief description; visitors arriving at the film's description from the Collection Spotlight were more likely to download it than visitors arriving from the less informative Pick List. Each of the five subsequent sharp drops can also be closely associated in time with an event involving the item. The first coincided with a referring link from the discussion site forums.somethingawful-.com and the second with a referring link from the extremely active weblog www.reason.com/hitandrun; in each case, the fraction of visitors arriving over these links who actually downloaded the film was very low. After the traffic from each of these referrers subsided, the batting average jumped back up. The final three drops correspond to technical failures on the site of varying lengths, which made it impossible to download the file.

Above we mentioned that a simpler alternative to HMMs, the computation of a sliding window of contiguous visits, is not effective for performing a comparable localization of events. The example we have been discussing here provides a good illustration of some of the difficulties. Perhaps the most common way of computing this type of sliding window is to convolve the 0–1-valued sequence $\mathbf{d} = (d_1, d_2, \ldots, d_T)$ with a Gaussian mask. In other words, letting $g(x)$ denote the Gaussian function $(1/\sqrt{2\pi})e^{-x^2/2\sigma^2}$, we create a smoothed sequence $\mathbf{d}' = (d'_1, d'_2, \ldots, d'_T)$, where $d'_i = \sum_{j=-k}^{k} g(j)d_{i+j}$ for a window size k. In this way, the smoothed quantity d'_i reflects the average of nearby elements of the original sequence $\mathbf{d}$, damped by the Gaussian multipliers.

In Fig. 3 we perform this computation with a Gaussian where $\sigma = 250$ and $k = 1{,}000$ (i.e., values are computed out to 4σ). Although the coarse shape of the plot resembles that of Fig. 2, the overall result is much noisier, and it is not clear how to localize particular discrete events. Larger values of σ produce plots that lose the overall shape as well, without becoming substantially less noisy.

The standard approach for identifying change points in such a smoothed signal would be to look for extreme points in the discrete analogue of the first derivative, $d'_{i+c} - d'_i$ for some constant $c > 0$, but this yields hundreds of such extrema for each item, a quantity that does not decrease significantly even with considerably more smoothing. To create a baseline for comparison with the HMM, we thus looked at just the set of extrema of largest absolute value in the discrete derivative, but we will show in the following section that this still does not perform nearly as well as the HMM at localizing the times of events involving individual items.

In summary, the example in this section suggests that an approach based on a HMM with a large number of underlying states can accurately localize points of discrete change and can capture changes in interest in an item over time scales that range from hours to months in duration.

Aligning Changes in Interest with External Events

The crux of the example in the previous section was that significant changes in the batting average for an active item are often correlated with "real-world events," both on and off the site, in which this item is featured. How general is this phenomenon? Here, we seek to address this question systematically, by studying the extent to which HMM state transitions can be aligned with events that occur nearby in time. In addition to providing an evaluation of our model's behavior, this type of alignment can have value for both users and administrators of the site. It is a way of identifying, from a large collection of

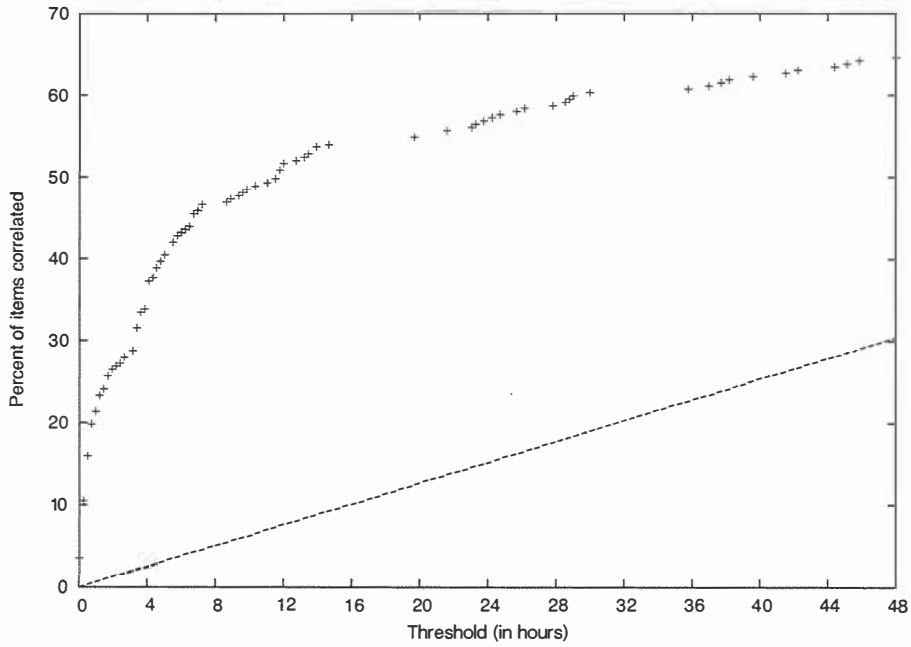


Fig. 4. As a function of a time threshold ±, the upper curve plots the size of the largest ±-hour alignment between major HMM transitions and observed events on the Internet Archive site, normalized by the total number of major transitions. The lower curve is a baseline for comparison: an upper bound on the corresponding quantity, with the set of major transitions replaced by a random set of points in time.

candidate events, those that demonstrated an impact on user behavior by substantially affecting the batting average of an item. Indeed, certain significant events that were not necessarily apparent at the time they occurred can be discovered after the fact through the effect that they have on items' batting averages.

Our approach is as follows. For each of the 100 most downloaded items on the Internet Archive, and for a period from September 2002 to May 2003, we compute all state transitions in the HMM defined in the previous section. We then discard those transitions in which the state changes only by the minimum increment of 0.01, as we are interested in detecting events that had a substantial impact on the batting average, whereas sequences of the minimum-size step of 0.01 occur when the batting average changes "smoothly" from one level to another without a big jump. We call the remaining transitions (257 over all 100 items) major transitions. We then check, for each of these major transitions, whether it occurred close to some observed event involving the item. To construct this set of observed events, we extract information from the Internet Archive's usage logs and a database that records all changes made by site administrators. Our full set of observed events is as follows:

- The appearance or disappearance of an item from a Collection Spotlight, Pick List, or top-level list of recent user reviews. (Each of these lists serves to highlight the item in a prominent location on the site.)
- The appearance of a link to an item's description from an active off-site referrer. We define this to be the first recorded visit from a referring URL that generated at least 100 visits total, with at least 25 visits occurring within 48 h of the first visit. Although these specific values are somewhat arbitrary, any similar values would achieve the goal of selecting off-site referrers that generated enough traffic to an item's description so as to have a potential impact on its batting average.
- The beginning or end of a technical failure on the site that prevented file downloads. These were determined by manual inspection of Internet Archive records and were assumed to involve all items.

There are a total of 1,978 events in this set, over all 100 items. Formally, we test for temporal proximity between major HMM transitions and observed events on the site as follows. We say that a δ-hour alignment between transitions and events is a collection of ordered pairs $(r_1, e_1), (r_2, e_2), \ldots, (r_k, e_k)$, where

(*i*) Each r_i is a major transition and the corresponding e_i is an event occurring at most ± hours away in time.
(*ii*) No transition or event occurs in more than one of the ordered pairs.

The effect of condition *i* is to require the transitions to localize events closely in time; the effect of condition *ii* is to prevent a single observed event from "explaining" multiple major transitions. Because the ideal is for major transitions to lie near observed events, we will say that such an alignment accounts for the transitions $r_1, r_2, \ldots r_k$. As a function of δ, the upper curve in Fig. 4 plots the size of the largest δ-hour alignment divided by the total number of major transitions. Thus we see, for example, that there is a 12-h alignment that accounts for roughly half (51.9%) of all major transitions.

To understand whether this is a significant overlap between transitions and events, we compare it to a random baseline. That is, approximately half of all major transitions can be accounted for by a 12-h alignment; is it likely that if we chose a random set of points in time, we could account for a comparable number? We address this question with the following calculation. There are 100 items under consideration, and we are focusing on a period of 260 days for each; so if we lay the time periods for each item end to end, we get an interval of 26,000 days, which is 624,000 h. Each of the 1,978 observed events "carves out" an interval of 2δ h in this timeline; so if we assume that none of these intervals overlap (which only helps the random baseline), then the probability a random point in time lies within δ hours of one of the observed events is at most $(2\delta)(1{,}978)/(624{,}000) \approx 0.00634\delta$. Thus, the expected fraction of random points that lie sufficiently close to an observed event is at most 0.00634δ, and hence the ratio of the largest ±-hour alignment to the total

number of random events can be at most this large. The lower curve in Fig. 4 shows a plot of this fraction as a function of δ. Thus we see, for example, that the maximum-size 12-h alignment will account for at most 7.6% of a random set of points in expectation, significantly less than the 51.9% obtained for the collection of major HMM transitions. Indeed, the probability of seeing a 12-h alignment account for 51.9% of a random sample of 257 points (the number of major transitions) is vanishingly small ($\approx 10^{-60}$).

Thus, there is significant overlap between events and transitions. Furthermore, it appears likely that many of the major transitions that went unexplained by observed events in fact have natural explanations that were not included in our event set. For example, when traffic from an active but short-lived referrer leads to a sharp change in the batting average, there is often a second major transition when this traffic dies down, but we do not generally have an observed event to relate this to. Second-order effects from referrers can be even harder to catch, as in a link from slashdot.org to the archive's top-level home page in February 2003 that drove a huge amount of traffic to the site. Certain items experienced a sudden change in batting average just after the appearance of this external link, because they were prominently featured on a top-level archive page and so a fair number of users arriving from slashdot.org went on to look at them, but no observed event was recorded for any of these items because the referring link was not directly to their description pages.

One conclusion from these missed explanations is in fact a promising one: that sharp changes in an item's batting average can often reveal genuine and significant events that are extremely hard to identify directly, even from extensive log data.

In the previous section, we discussed the difficulties with localizing observed events using change points in a Gaussian-smoothed version of the 0-1 sequence of download decisions. Continuing the notation used in that discussion, we took the set of times at which the 257 largest absolute values occurred in the discrete derivative $d_{i+c}^{l} - d_{i}^{l}$ with $c = 6$, and computed the largest δ-hour alignment of this set with the observed events as a function of δ. (Empirically, we found that the choice of $c = 6$ produced the most favorable results for this method, and the choice of the top 257 changes was made so as to produce a set of the same size as the collection of all major HMM transitions.) Although this alignment outperformed the random baseline discussed above, it was significantly smaller than the corresponding alignment of major HMM transitions with observed events, across all values of δ. For example, the largest 12-h alignment in the case of Gaussian smoothing accounted for 19.5% of all points, compared with 7.6% for a random set and 51.9% for major HMM transitions.

As a final point of discussion, we note that we would get much smaller numbers if we studied the converse question: what fraction of observed events in our set occur close in time to a major HMM transition? The point is that although we expect major transitions to align with some observed event, we do not necessarily expect each observed event to correlate with a corresponding change in the HMM. This idea is consistent with an issue addressed earlier, that certain observed events have a measurable impact on an item's batting average, but many do not. Discrete changes in the batting average over time can thus be useful in identifying the extent to which particular links and other forms of highlighting did or did not have an impact on an item's popularity.

Related Work

Much of the prior work on usage data has addressed the problem of collaborative filtering, recommending items to users based on their pattern of past behavior on the site. Research on collaborative filtering has developed approaches that build models of individual users, so as to predict a user's interest in items (8), as well as approaches that build models of item-to-item similarity aggregated over many users, as one sees at sites like amazon.com (9).

More closely related to the issues we consider here is recent work on predicting purchase conversion on an e-commerce site, estimating the probability that an individual user will perform a purchase based on his or her browsing pattern (e.g., refs. 10–12). One key respect in which the work on purchase conversion differs from our approach is in its emphasis on a per-user style of analysis, focusing on a single user's behavior across many items, as opposed to the per-item analysis we undertake here, which considers the behavior of many users in response to a single item.

There has also been work on usage data in the context of information visualization, helping users explore a site by revealing the collective behavior of other users (13, 14). Our analysis of events and their correlation with changes in batting averages offers a way to summarize collective user behavior from a very different perspective, and it would be interesting to see how far these approaches could be integrated.

Finally, it is interesting to note that the success of probabilistic models with explicit state discussed above, compared with algorithms based on local averaging, follows a closely analogous theme in computer vision, where Markov random field models have gained prominence as a technique for dealing with discontinuities in images (15). Our approach here is also motivated by the use of state transitions to model discrete "bursts" in online event sequences (16).

Further Directions

A site as active as the Internet Archive has events of many different kinds impinging on it simultaneously: users view and download items, write reviews, and post messages to discussion boards; active external sites discuss the archive and drive traffic to it; world events generate interest in particular items at the archive. Our probabilistic model for identifying changes in the batting average allows us to analyze one of these streams of actions, the sequence of download decisions, in a principled fashion. Our evaluation in the previous section represents a step toward the simultaneous analysis of multiple streams of events, through the alignment of events on the archive site with discrete changes in the batting average.

It will be interesting to carry this style of analysis further. For example, although we have informally discussed the notion that an external event such as an active referrer may "cause" a change in the batting average, we have refrained from trying to make the concept of causality precise; thus our evaluation above focused purely on proximity in time between events and transitions. Quantifying the notion of causality more precisely is a very natural open question and, from our experience with the data, a difficult one, given the large number of factors that concurrently influence user interest in an item, and the difficulty in isolating the contribution of these factors separately.

Finally, it is clear that tracking just the description-to-acquisition behavior of users has already exposed a rich pattern of activity that varies across time and subpopulations. But it would also be valuable to look at more extensive representations of user behavior, by tracking longer user paths through the site; this offers the chance to make richer inferences about both group and individual user intentions (10–12, 17–20), although it becomes correspondingly harder to interpret the usage data. Ultimately, by considering an increasing level of detail in the dynamics of traffic at an active web site, we can hope to achieve more detailed insight into the collective behavior of the crowds that congregate there.

We thank Brewster Kahle for his insights and support throughout the course of this work. This work has been supported in part by a David and Lucile Packard Foundation Fellowship and National Science Foundation Information Technology Research Grant IIS-0081334.

1. Chakrabarti, S. (2002) *Mining the Web* (Morgan Kaufman, San Francisco).
2. Ginsparg, P. (1996) in *Electronic Publishing in Science*, eds. Shaw, D. & Moore, H. (United Nations Educational, Scientific, and Cultural Organization, Paris), pp. 83–88.
3. Lawrence, S., Giles, C. L. & Bollacker, K. (1999) *IEEE Computer* **32,** 67–71.
4. Mitzenmacher, M. (2003) *Internet Math.*, in press.
5. Rabiner, L. (1989) *Proc. IEEE* 77, 257–286.
6. Felzenszwalb, P. & Huttenlocher, D. (2000) in *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference*, eds. Kriegman, D. & Forsyth, D. (IEEE Computer Society Press, Los Alamitos, CA), pp. 66–73.
7. Felzenszwalb, P., Huttenlocher, D. & Kleinberg, J. (2003) *Adv. Neural Information Processing*, in press.
8. Resnick, P. & Varian, H. (1997) *Commun. ACM* **40,** 56–58.
9. Linden, G., Smith, B. & York, J. (2003) *IEEE Internet Computing* **7,** 76–80.
10. Moe, W. & Fader, P. (2000) *University of Pennsylvania Wharton School Marketing Working Paper 00-023* (Univ. of Pennsylvania, Philadelphia).
11. Montgomery, A., Li, S., Srinivasan, K. & Liechty, J. (2003) *Carnegie Mellon Graduate School of Industrial Administration Working Paper 2003-E26* (Carnegie Mellon, Pittsburgh).
12. Sismeiro, C. & Bucklin, R. (2002) *UCLA Anderson School Working Paper 2002* (Univ. of California, Los Angeles).
13. Eick, S. (2001) *Commun. ACM* **44,** 45–50.
14. Fry, B. (2000) Master's thesis (Massachusetts Institute of Technology, Cambridge).
15. Li, S. (1995) *Markov Random Field Modeling in Computer Vision* (Springer, Berlin).
16. Kleinberg, J. (2002) in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, eds. Hand, D., Keim, D. & Ng, R. (Assoc. Computing Machinery, New York), pp. 91–101.
17. Bucklin, R., Lattin, J., Ansari, A., Gupta, S., Bell, D., Coupey, E., Little, J., Mela, C., Montgomery, A. & Steckel, J. (2002) *Marketing Lett.* **13,** 245–258.
18. Cadez, I., Heckerman, D., Meek, C., Smyth, P. & White, S. (2003) in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, eds. Simoff, S. & Zaiane, O. (Assoc. Computing Machinery, New York), pp. 280–284.
19. Heer, J. & Chi, E. (2002) in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, eds. Torveen, L. & Wixon, D. (Assoc. Computing Machinery, New York), pp. 243–250.
20. Huberman, B., Pirolli, P., Pitkow, J. & Lukose, R. (1998) *Science* **280,** 95–97.

Evolution of document networks

Filippo Menczer†

School of Informatics, Indiana University, Bloomington, IN 47408

How does a network of documents grow without centralized control? This question is becoming crucial as we try to explain the emergent scale-free topology of the World Wide Web and use link analysis to identify important information resources. Existing models of growing information networks have focused on the structure of links but neglected the content of nodes. Here I show that the current models fail to reproduce a critical characteristic of information networks, namely the distribution of textual similarity among linked documents. I propose a more realistic model that generates links by using both popularity and content. This model yields remarkably accurate predictions of both degree and similarity distributions in networks of web pages and scientific literature.

There are important social and economic implications in explaining how the topology evolves in information networks such as the World Wide Web. Although text analysis has been used for a long time to analyze documents, extract their meaning, retrieve information, and map knowledge domains (1–3), link analysis is increasingly used by search engines and digital libraries to estimate the importance or reputation of documents (4–6) and to map documents into topical clusters (7–10).

A number of models have been proposed to explain the growth of complex networks exhibiting characteristics such as the small-world property (high clustering and low diameter) and the scale-free property (power-law distribution of degree). A few representative models are reviewed in *Background*. Which of these competing theories is more plausible? In *Validating Prior Models*, I show that the answer is none: Although they all correctly predict degree distributions, they fail at predicting another observable feature of the information network, namely the distribution of textual similarity among linked documents. In *Degree-Similarity Mixture Model*, I propose a growth model to explicitly capture the trade-off between an author's desire to link related and popular documents. The model is validated against two data sets: a network of web pages sampled from the Open Directory Project (DMOZ; http://dmoz.org) and a collection of scientific articles published in PNAS. Web pages are connected by hyperlinks and articles by citations. Numerical simulations are used to generate predictions of degree and similarity distributions for both data sets.

Background

Since the discovery of scale-free and small-world phenomena in web pages (11–15) and bibliographic collections (16–18), physicists and computer scientists have developed growth theories that model the behavior of authors linking new pages to explain the emergence of these critical network properties (9, 19). Most growth models are based on some form of preferential attachment, whereby one node at a time is added to the network with new edges to existing nodes selected according to some probability distribution.

In the best known preferential attachment model, a node i receives a new edge with probability proportional to its current degree, $\Pr(i) \propto k(i)$ (13). This so-called BA model generates networks with power-law degree distributions, in which the oldest nodes are those with highest degree. The copying model and its extensions implement equivalent rich-get-richer processes based on local walks, without requiring explicit knowledge of degree (20–22).

To give newer nodes a chance to compete for links, an extension of the preferential attachment model is based on linking to a node dependent on its degree with some probability or to a uniformly chosen node with the remaining probability (23, 24). Such a mixture model generates networks that can fit the power-law degree distribution of the entire web as well as the different distributions observed in subsets of the web such as university and business homepages (25).

Some theories have explored similar mixture models in which links are created according to a trade-off between graph degree and metric distance measures, showing that certain trade-off regimes lead to power-law degree distributions (26, 27).

To study the decision process by which authors link documents, let us consider the relationship between the probability that two documents are linked and their content (text) similarity. Content similarity can be measured by the cosine metric traditionally used in information retrieval (1):

$$\sigma_c(d_1, d_2) = \frac{\|\vec{d}_1 \cdot \vec{d}_2\|}{\|\vec{d}_1\| \cdot \|\vec{d}_2\|}, \qquad [1]$$

where $\vec{d}$ is a vector space representation of the text in document d. Link probability can be approximated by a link similarity (or neighborhood) metric defined as the Jaccard coefficient:

$$\sigma_l(d_1, d_2) = \frac{|U_{d_1} \cap U_{d_2}|}{|U_{d_1} \cup U_{d_2}|}, \qquad [2]$$

where U_d is the set representing d's neighborhood, which consists of inlinks and outlinks for web pages and citations and references for articles (in which case σ_l is akin to cocitation and bibliographic coupling). To visualize the relationship between text and link similarity, one can map the joint distribution of σ_c and σ_l across pairs of documents. Fig. 1 shows two such maps, obtained from collections of web pages and PNAS articles. These maps demonstrate that for web pages as well as scientific papers, authors connect documents in a way that is significantly (although not strongly) correlated with the similarity of those documents to their own. The Pearson correlation coefficient between σ_c and σ_l is 0.10 for web pages (3.8×10^9 pairs) and 0.12 for PNAS articles (7.5×10^6 pairs).

To quantify the dependence of the web's link topology on content, I considered the conditional probability that the link neighborhood between two web pages is above some threshold λ, given that the two pages have some content similarity κ, as a function of κ:

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

†E-mail: fil@indiana.edu.

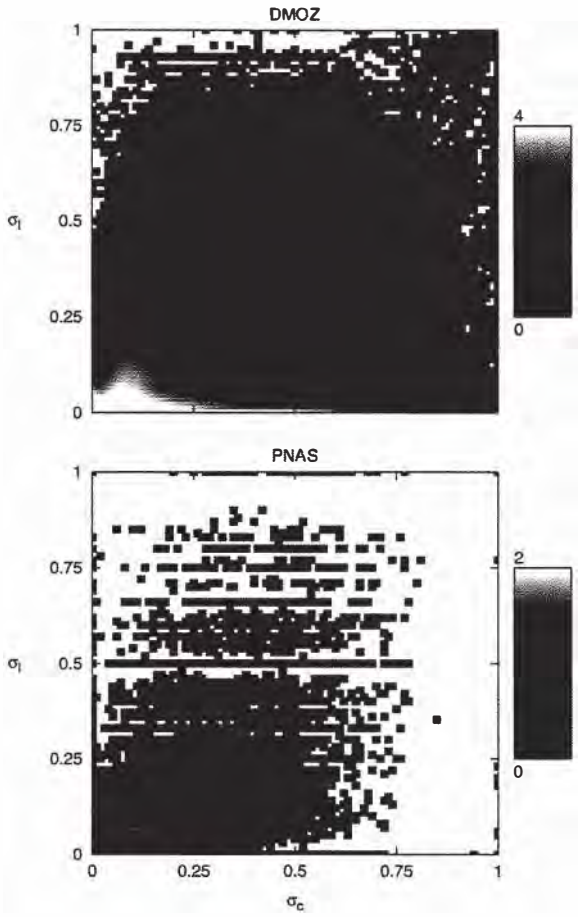


Fig. 1. Joint distribution maps for content and link similarity across pairs of web pages (*Upper*) and scientific articles (*Lower*). Colors code the $\log_{10}$ of the number of pairs with the corresponding similarity coordinates. The web data are based on a stratified sample of 109,648 pages from the DMOZ, with their inlinks and outlinks, crawled in 2002. The article data are based on the titles, abstracts, and references of 15,785 articles published in PNAS between 1997 and 2002.

$$\Pr(\lambda|\kappa) = \frac{|(p,q):\sigma_c(p,q) = \kappa \wedge \sigma_l(p,q) > \lambda|}{|(p,q):\sigma_c(p,q) = \kappa|}, \qquad [3]$$

where p, q are two web pages. An interesting phase transition was observed between two distinct regions around a critical distance κ^* independent of λ (28). For $\kappa > \kappa^*$, the probability that two pages are neighbors does not seem to depend on their content similarity; for $\kappa < \kappa^*$, the probability decreases according to a power-law $\Pr(\lambda|\kappa) \sim \kappa^{\gamma}$, with the decay exponent γ growing linearly with λ. This observation suggested a content-based growth model in which an author tends to link a new page to the most popular among related (similar) pages and with decreasing probability to less similar ones. As in the BA model, at each step t one new page t is added, and m new links are created from t to m existing pages, each selected from $\{i, i < t\}$ with probability:

$$\Pr(i,t) = \begin{cases} \dfrac{k(i)}{mt} & \text{if } \sigma_c(i,t) > \kappa^* \\ c\sigma_c^{\gamma}(i,t) & \text{otherwise} \end{cases}, \qquad [4]$$

where m, κ^*, and γ are constants derived from the data and c is a normalization factor. This degree-similarity phase model accurately predicted the degree distribution of the web pages in the sample from which the data were derived and was the first model to do so taking content into account (28).

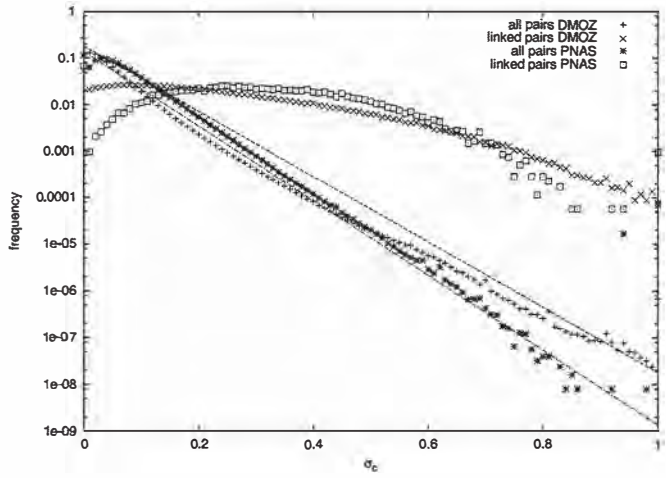


Fig. 2. Content similarity distributions for web pages (DMOZ) and scientific articles (PNAS). The distributions across linked documents are clearly different from the background distributions. The latter are exponential for both data sets (exponential fit curves appear as lines in the log-linear plot).

Validating Prior Models

Given that all of the models described in *Background* can predict the degree distribution of web pages and scientific articles, which is most plausible and/or powerful in explaining the emerging topology of information networks? To answer this question, we need an independent observation from the data, in addition to degree. I turned to the distribution of content similarity between linked (neighbor) pages. Fig. 2 demonstrates that this distribution is qualitatively different from the background similarity distribution for both web pages and scientific articles, clearly indicating that content must play a role in the evolution of information networks. I then looked for a model capable of predicting both the degree distributions and the similarity distributions among linked documents. Such a model would be more plausible than models predicting degree alone.

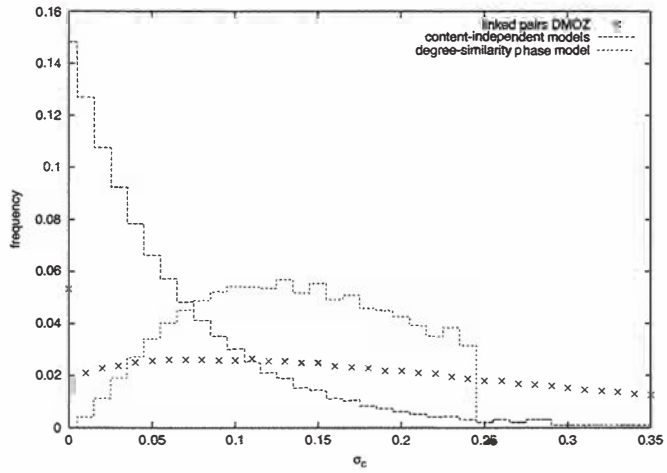


Fig. 3. Content similarity distributions across linked pages generated by simulating various growth models for web pages, compared with DMOZ data. In all simulations the parameters are set to match or fit the DMOZ data: n = 109,648 nodes, m = 15 links, and, in the degree-similarity phase simulation, κ^* = 0.25 and γ = 1.7.

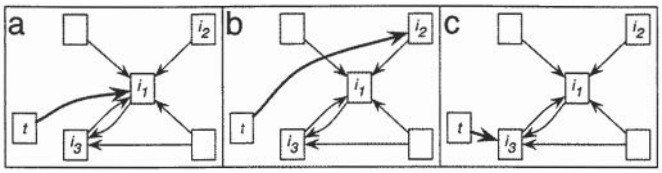


Fig. 4. (a) In all preferential attachment models, a new page t is likely to be linked to a page with high degrees such as i_1. (*b*) In the degree-uniform mixture model, there is also some probability to link to any random page such as i_2. (*c*) In the degree-similarity mixture model, t is more likely to be linked to a page that has high degree but is close in content space, i.e., similar to t, such as i_3.

The models in *Background* were validated by numerical simulations, with content similarity drawn from the exponential distributions obtained by fitting the background distributions in Fig. 2: $\Pr(\sigma_c) \sim 10^{-\mu\sigma_c}$ where $\mu = 7$ for web pages and $\mu = 8$ for PNAS articles. The degree distribution of the data are well matched by those generated by the mixture model (25) and the degree-similarity phase model (28). On the other hand, as shown in Fig. 3, for web pages the growth models that do not consider content (13, 20, 25) predictably generate similarity distributions across linked pages that mirror the background exponential distribution. The degree-similarity phase model does somewhat better, but it generates a distribution that goes to zero too rapidly for small and large σ_c. Results are analogous for PNAS articles. Thus, none of the growth models outlined above generates distributions of textual similarity across linked documents in qualitative agreement with the data.

Degree-Similarity Mixture Model

The class of mixture models has a free parameter that can be tuned to fit the data. At each step, one new document is added, and m new links or references are created from it to existing documents. At time t the probability that the ith document is selected and linked from the tth document is

$$\Pr(i) = \alpha \frac{k(i)}{mt} + (1 - \alpha)\overline{\Pr}(i)\,, \qquad [5]$$

where $i < t$ and $\alpha \in [0,1]$ is a preferential attachment parameter (Fig. 4*a*). In the degree-uniform mixture model (25) $\overline{\Pr}(i) = 1/t$, the uniform distribution (Fig. 4*b*). Let us now introduce an alternative degree-similarity mixture model in which

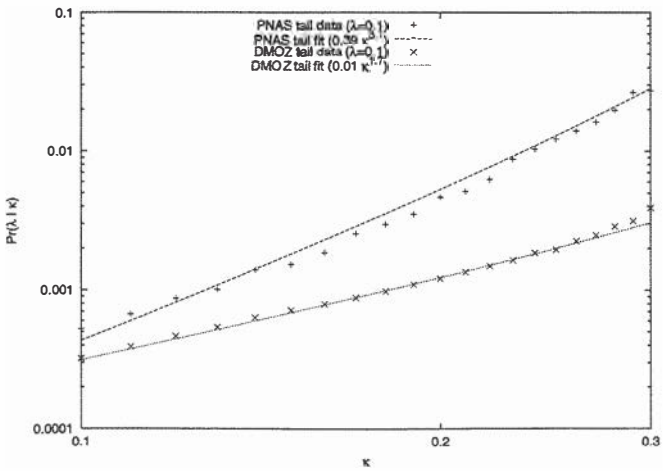


Fig. 5. Tails of conditional link probability $\Pr(\lambda|\kappa)$ as a function of κ for pairs of web pages (DMOZ) and PNAS articles, with power-law fit exponents $\gamma = 1.7$ and $\gamma = 3.1$ respectively.

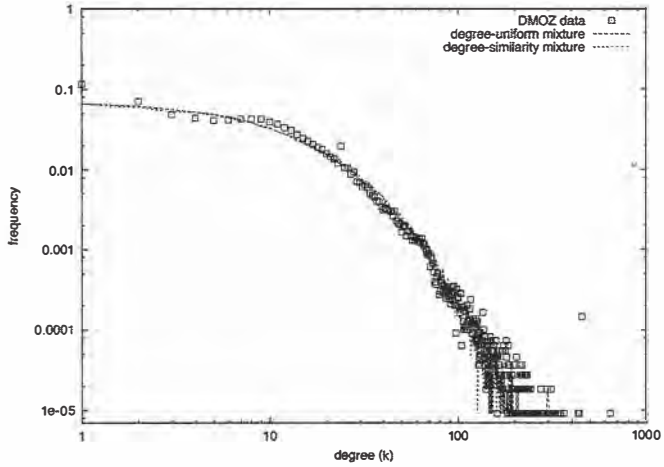


Fig. 6. Degree distributions of web pages predicted by simulating the two mixture models. In the degree-uniform mixture simulation, $\alpha = 0.3$; in the degree-similarity simulation, $\alpha = 0.2$ and $\gamma = 1.7$. All parameters were set by fitting the DMOZ data.

$$\overline{\Pr}(i) \propto \left(\frac{1}{\sigma_c(i,t)} - 1\right)^{-\gamma}, \qquad [6]$$

where γ is a constant (Fig. 4*c*). Like the degree-similarity phase model, this model is inspired by the idea that authors tend to link new documents to popular and related ones and by the observation that link probability between two web pages decays with decreasing similarity as a power-law $\Pr(\lambda|\kappa) \sim \kappa^{\gamma}$ with $\gamma = 1.7$ for $\lambda = 0.1$ (Fig. 5). However, the free parameter α in the degree-similarity mixture allows us to explicitly model the trade-off between linking to related (similar) versus popular (high-degree) documents.

To validate the degree-similarity mixture model, the networks of web pages and PNAS articles were built by simulation and compared with those obtained by simulating the degree-uniform mixture model. Figs. 6 and 7 show the predictions generated for web pages. Although both models accurately predict the degree distribution, only the degree-similarity mixture model reasonably approximates the similarity distribution.

The PNAS article data were analyzed analogously to the DMOZ data, yielding a conditional citation probability with a tail that scales as a power-law $\Pr(\lambda = 0.1|\kappa) \sim \kappa^{\gamma}$ just as for web

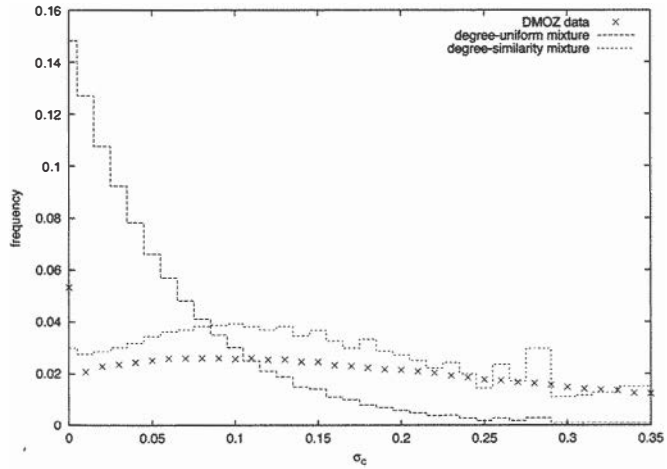


Fig. 7. Distribution of content similarity among linked web pages predicted by simulating the two mixture models.

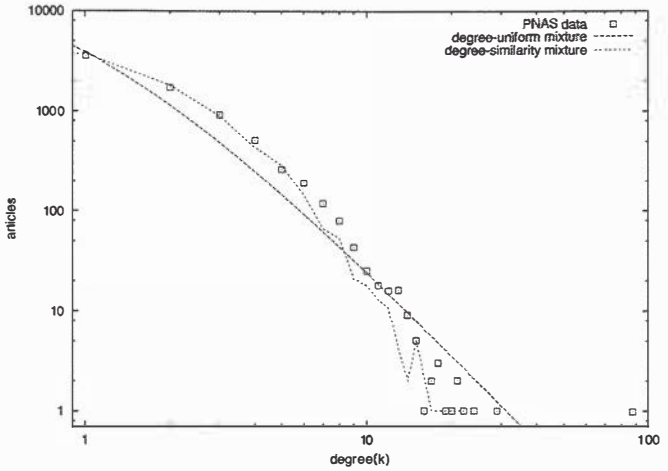


Fig. 8. Distributions of degree (citation count) of PNAS articles, as predicted by the two mixture models. In all of the simulations, $n = 15{,}785$ nodes and $m = 1$ reference. In the degree-uniform mixture simulation, $\alpha = 0.5$; in the degree-similarity simulation, $\alpha = 0.1$ and $\gamma = 3.1$. All parameters were set by matching or fitting the PNAS data (only references to other papers in the PNAS collection were considered).

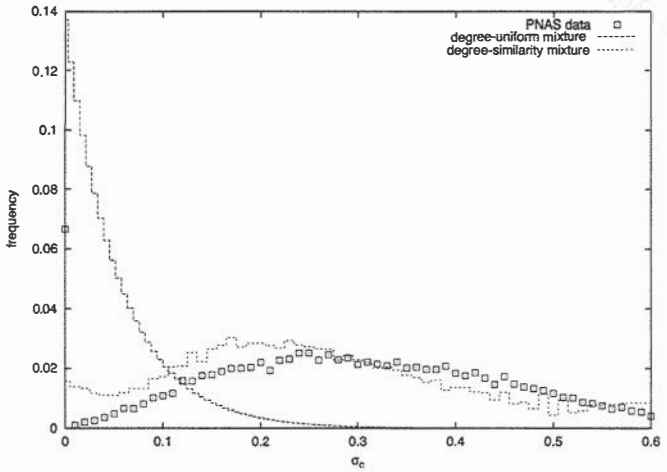


Fig. 9. Distribution of content similarity among titles and abstracts of articles that cite one another, as predicted by the two mixture models.

pages, but with a larger exponent $\gamma = 3.1$ (Fig. 5). Figs. 8 and 9 show the predictions generated by simulating the growth of the PNAS article network according to the two mixture models. Both models accurately predict the distribution of citation counts, although the degree-similarity model fits the data better. Again, the degree-similarity mixture model generates a similarity distribution in remarkable agreement with the data.

Conclusion

In this paper I have shown that existing growth models for document networks generate the wrong predictions for the distribution of content similarity across linked documents. Models that do not take content into account yield distributions that are heavily skewed toward low similarities because those are exponentially more frequent in the data. On the contrary, if authors tend to link and cite related documents, one would expect a similarity distribution with a fatter tail and a peak for $\sigma_c > 0$, as displayed by both web pages and scientific articles. This behavior is captured by the degree-similarity mixture model.

The similarity measures and document representations used in the presented analysis and model are quite crude. Cosine similarity here is based on simple term frequencies. The Jaccard coefficient for link similarity not only is a rough approximation of link probability for web pages but is further limited by incomplete knowledge owing to the necessary reliance on search engines for inlink information. One natural direction for future work is to repeat the analysis and validate the degree-similarity mixture model by using more sophisticated document representations and similarity measures (2, 3, 29, 30). Another is to extend the validation of the model to see whether it can predict additional properties of the networks, such as clustering coefficient and degree correlation (18, 22). Finally, further insight must be gained by studying the relationship between the mechanism studied here (linking similar documents) and other processes likely to play a role in the evolution of link/citation networks, such as copying (22) and coauthorship (17, 18).

The results presented here strongly suggest that page content cannot be neglected when we try to understand the evolution of document networks. The tension between referring to popular versus related documents provides us with a plausible and unified model of how authors link nodes in such different networks as the web and the scientific literature. This model generates remarkably accurate predictions of how such a process can lead to the emergent link and content structure of document spaces.

I thank Jon Kleinberg, Rob Axtell, David Aldous, Lászlo Barabási, Reka Albert, Mark Newman, Lada Adamic, Alessandro Vespignani, and Katy Börner for reviewing an earlier draft of this manuscript; Rich Shiffrin and two anonymous reviewers for providing helpful suggestions; the Open Directory Project for the DMOZ data; and the National Academy of Sciences for the PNAS data. This work was supported by National Science Foundation Career Award IIS-0133124/0348940.

1. Salton, G. & McGill, M. (1983) *An Introduction to Modern Information Retrieval* (McGraw–Hill, New York).
2. Belew, R. (2000) *Finding Out About: A Cognitive Perspective on Search Engines and the WWW* (Cambridge Univ. Press, Cambridge, U.K.).
3. Börner, K., Chen, C. & Boyack, K. (2003) *Annu. Rev. Inf. Sci. Technol.* **37,** 179–255.
4. Brin, S. & Page, L. (1998) *Comput. Networks ISDN Syst.* **30,** 107–117.
5. Kleinberg, J. (1999) *J. Assoc. Comput. Mach.* **46,** 604–632.
6. Mendelzon, A. & Rafiei, D. (2000) *IEEE Data Eng. Bull.* **23,** 9–16.
7. Kleinberg, J. & Lawrence, S. (2001) *Science* **294,** 1849–1850.
8. Girvan, M. & Newman, M. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 8271–8276.
9. Henzinger, M. & Lawrence, S. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5186–5191.
10. Hopcroft, J., Khan, O., Kulis, B. & Selman, B. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5249–5253.
11. Albert, R., Jeong, H. & Barabási, A.-L. (1999) *Nature* **401,** 130–131.
12. Huberman, B. & Adamic, L. (1999) *Nature* **401,** 131.
13. Barabási, A.-L. & Albert, R. (1999) *Science* **286,** 509–512.
14. Broder, A., Kumar, S., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000) *Comput. Networks ISDN Syst.* **33,** 309–320.
15. Adamic, L. & Huberman, B. (2000) *Science* **287,** 2115.
16. de Solla Price, D. (1965) *Science* **149,** 510–515.
17. Newman, M. E. J. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5200–5205.
18. Börner, K., Maru, J. T. & Goldstone, R. L. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5266–5273.
19. Dorogovtsev, S. & Mendes, J. (2003) *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford Univ. Press, Oxford).
20. Kleinberg, J., Kumar, S., Raghavan, P., Rajagopalan, S. & Tomkins, A. (1999) *Lect. Notes Comput. Sci.* **1627,** 1–18.
21. Kumar, S., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A. & Upfal, E. (2000) in *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science* (IEEE Comput. Soc., Silver Spring, MD), pp. 57–65.
22. Vazquez, A. (2003) *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **67,** 056104.

23. Dorogovtsev, S., Mendes, J. & Samukhin, A. (2000) *Phys. Rev. Lett.* **85,** 4633–4636.
24. Cooper, C. & Frieze, A. (2001) in *Proceedings of the 9th Annual European Symposium on Algorithms*, Lecture Notes in Computer Science, ed. Meyer auf der Heide, F. (Springer, Berlin), Vol. 2161, pp. 500–511.
25. Pennock, D., Flake, G., Lawrence, S., Glover, E. & Giles, C. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 5207–5211.
26. Fabrikant, A., Koutsoupias, E. & Papadimitriou, C. (2002) in *Proceedings of the 29th International Colloquium on Automata, Languages, and Programming*, Lecture Notes in Computer Science, eds. Widmayer, P., Triguero, F., Morales, R., Hennessy, M., Eidenbenz, S. & Conejo, R. (Springer, Berlin), Vol. 2380, pp. 110–122.
27. Aldous, D. (2003) arXiv.:cond-mat/0304701
28. Menczer, F. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 14014–14019.
29. Ganesan, P., Garcia-Molina, H. & Widom, J. (2003) *Assoc. Comput. Mach. Trans. Inf. Syst.* **21,** 64–93.
30. Landauer, T. K., Laham, D. & Derr, M. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5214–5219.

The simultaneous evolution of author and paper networks

Katy Börner[†‡], Jeegar T. Maru[§], and Robert L. Goldstone[¶]

[†]School of Library and Information Science and Departments of [§]Computer Science and [¶]Psychology, Indiana University, Bloomington, IN 47405

There has been a long history of research into the structure and evolution of mankind's scientific endeavor. However, recent progress in applying the tools of science to understand science itself has been unprecedented because only recently has there been access to high-volume and high-quality data sets of scientific output (e.g., publications, patents, grants) and computers and algorithms capable of handling this enormous stream of data. This article reviews major work on models that aim to capture and recreate the structure and dynamics of scientific evolution. We then introduce a general process model that simultaneously grows coauthor and paper citation networks. The statistical and dynamic properties of the networks generated by this model are validated against a 20-year data set of articles published in PNAS. Systematic deviations from a power law distribution of citations to papers are well fit by a model that incorporates a partitioning of authors and papers into topics, a bias for authors to cite recent papers, and a tendency for authors to cite papers cited by papers that they have read. In this TARL model (for topics, aging, and recursive linking), the number of topics is linearly related to the clustering coefficient of the simulated paper citation network.

Models capturing the structure and evolution of mankind's scientific endeavor are expected to provide insights into the inner workings of science. They are developed to provide objective guidance to augment decisions concerning resource allocation (identification of research frontiers, determining award amount, many small vs. a few large grants), optimum interdisciplinary collaboration (too little collaboration might lead to duplication, too much may lead to rather shallow science), the influence of publishing mechanisms (books vs. fast e-journals), and so on.

Two kinds of models are commonly distinguished: descriptive models that aim to describe the major features of a (typically static) data set and process models that model the mechanisms and temporal dynamics by which real-world networks (e.g., coauthor or paper citation networks) are created. Most research in bibliometrics (1), scientometrics (2), or knowledge domain visualizations (3) has focused on descriptive models. For example, research has studied the statistical patterns of coauthorship networks, paper citation networks, individual differences in citation practice, the composition of knowledge domains, and the identification of research fronts as indicated by new but highly cited papers. Recent work in statistical physics and sociology aims to design process models. Of particular interest is the identification of elementary mechanisms that lead to the emergence of small-world (4, 5) and scale-free network structures (6, 7).

The model proposed in this article is unique in that it simulates the simultaneous growth of more than one network structure, here authors and papers. The core assumption is that the twin networks of scientific researchers and scholarly articles mutually support one another. Researchers connect articles to one another in cocitation networks, and articles link researchers to one another in coauthorship networks.

The model provides a grounded mechanism for modeling the "rich-get-richer" phenomenon for paper citation networks as an emergent property of the elementary networking activity of authors reading and citing articles and also the references listed in read articles. The generalized rich-get-richer phenomenon is also known as the Mathew effect (8), cumulative advantage (9), or preferential attachment (10).

The growth of scientific publications and citations is governed by two underlying processes: growth and aging (11). Growth seems to be important for the development of scale-free networks. Aging is an antagonistic force to preferential attachment. Even highly connected nodes typically stop receiving links after time has passed. The bias to cite newer papers frequently prevents a scale-free distribution of connectivity (12). In the proposed model, an aging bias offsets the rich-get-richer phenomenon for paper citation networks.

A 20-year data set of articles published in PNAS is used to validate the model in terms of major network properties of the interlinked coauthor and paper citation networks.

The subsequent sections review related research on descriptive and process models of coauthor and paper citation networks, discuss desirable features and basic assumptions of the process model, validate the model by comparing simulated data to a 20-year ISI[‖] PNAS data set, and discuss the influence of model parameters such as the aging of papers in terms of their power to attract citations, the number of topics, and the length of the chain of references that authors consider when making citations. The article concludes with a discussion and outlook.

Related Work

There is a long history in bibliometrics (1) and scientometrics (2) of describing the structure and evolution of science (3). As early as 1964, Garfield and his colleagues (13) proposed using citation data to study and write the history of science. Citation data has also been used to identify the associations between authors, publications, patents, grants, data, and more recently genes, proteins, diseases, etc. Associations have been discovered over time, space, and fields to identify changing frontiers of science (14), measure science (15), or recognize research fronts (16).

Research on process models seeks to simulate, statistically describe, or formally reproduce statistical characteristics of interest. Of particular interest are models that "conform to the measured data not only on the level where the discovery was

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

[‡]To whom correspondence should be addressed. E-mail: katy@indiana.edu.

[‖]These data are extracted from *Science Citation Index Expanded* [Institute for Scientific Information, Inc. (ISI), Philadelphia, PA; Copyright ISI]. All rights reserved. No portion of these data may be reproduced or transmitted in any form or by any means without the prior written permission of ISI.

www.pnas.org/cgi/doi/10.1073/pnas.0307625100

originally made but also at the level where the more elementary mechanisms are observable and verifiable" (17).

Recent work in statistical physics aims to design models and tools to analyze the statistical mechanics of topology and dynamics of diverse physical, biological, and social networks. A major goal is to identify elementary mechanisms that lead to the emergence of small-world (4, 5) and scale-free network structures (6, 7) commonly observed in the real world.

Small-world networks have a short average path length among nodes but a high local clustering coefficient compared to random networks (18). Important small-world graph properties include the number of vertices (n), the average degree $\langle k \rangle$, the characteristic path length (l), and the clustering coefficient (C). The degree of a vertex is the total number of its connections. The characteristic path length describes how far apart any two nodes in the graph network are. It is computed by determining the shortest path l (i,j) between any two nodes i and j in the network and calculating the average of all l. The clustering coefficient is a more local measure of how "cliquish" a graph is or how tightly nodes in the graph are connected to each other. If a node has K edges that connect it to its neighbors, then the node's clustering coefficient is given by $C = N/(K^*(K - 1)/2)$, where N is the number of edges connecting neighbors of the node to each other. The strength of a connection (e.g., the number of times two authors wrote papers together) is not considered during the computation of l or C.

In scale-free networks, the frequency f of the degree of connectivity k of a vertex is a power function of k, $f \approx k^{-}\gamma$. Examples of real-world data sets that are well approximated by power law relationships include actor collaborations, power grids, and the worldwide web (10). For these data sets, the power law applies over many orders of magnitude, and hence these networks are known as "scale-free." With very few highly interlinked nodes and many weakly interlinked nodes, scale-free networks are surprisingly robust against random deletion of edges, e.g., network failures (19).

The Watts–Strogatz model was the first model to generate graphs with small-world properties (20). Their process model starts with a regular lattice network configuration. Each edge is redirected with a probability p to another randomly selected node. This process model is of limited direct utility for coauthor and paper citation networks because in those networks the links are fixed; no rewiring takes place. That is, once a paper has cited another paper, or two authors have collaborated on a paper, these associations are forever part of the permanent historical record.

The Barabási and Albert (BA) model has been a highly influential and successful attempt to simulate networks that show scale-free properties (10). It starts with a small number (N_0) of nodes, and at every time step, a new node is added as well as a set of m new edges that link the new node to the nodes already present in the system. The probability p that a new node will be connected to node i is proportional to the degree k_i of node i. Hence a new node is preferentially attached to an already highly connected node. After t time steps the network has $n = t + N_0$ nodes and m^*t edges. This network evolves into a stationary scale-free state with the probability that a node has k edges following a power law with an exponent $\gamma_{BA} = 3$. Gradually adding nodes to the network over time appears to be critical in obtaining scale-free distributions (21).

Copying behavior was introduced as an alternative to explain the power law degree distribution for the worldwide web (22). Recent work[††] models the probability distribution of citations by copying references used in other papers. The resulting network quantitatively matches the citation distribution observed in real citation networks. Vazquez[‡‡] even suggested that authors do a recursive bibliographic search. In his model a new node is connected to a randomly selected node as well as nodes linked from (referenced by) this node. Although these models attempt to capture preferential attachment, less is known about their small-world properties. Numerous other attempts to model small-world and scale-free networks are reviewed in refs. 4 and 7.

A number of mathematical models of network evolution have been developed in sociology. Several models (25) assume a fixed number of edges. Snijders (26) proposed a class of statistical models for longitudinal network data that assumes a directed graph with a fixed set of actors. However, neither the number of nodes nor edges is fixed for evolving coauthor or paper citation networks. The model by Gilbert (27) aims to simulate the structure of academic science. It assumes that papers generate future papers, giving authors a rather incidental role. The model was validated based on the number and distribution of citation counts. The small-world and scale-free properties of the resulting networks are unknown.

To our knowledge no algorithmic approach exists that simultaneously models the evolution of different networks such as coauthor and paper citation networks within an ecology of multiple interacting networks. Here, we argue that to fully understand the structure, evolution, and utilization of networks, coauthor and paper citation networks need to be considered simultaneously. For example, to understand how knowledge diffuses across authors via their papers at the same time that new authors and papers are accumulated, it is essential to model the coupled growth of both network structures.

Process Model for Author–Paper Networks

This section motivates the features and simplifying assumptions of a process model for the simultaneous growth of coauthor and paper citation networks as seen in citation databases like PNAS. Given the importance of the interplay of topics, aging, and recursive follow-up of links (here citation references), it was named the TARL (topics, aging, and recursive linking) model.

The TARL model attempts to capture the roles of authors and papers in the production, storage, and dissemination of knowledge. Information diffusion is assumed to occur directly via coauthorships and indirectly via the consumption of other authors' papers. It assumes the existence of a set of authors and papers. Each author and paper is assigned a single topic. Ideally, several levels of topics would be organized hierarchically in terms of specificity. The same paper may belong to the coarse topic of immunology, the more specific topic of HIV infection, and the still more specific topic of hemolytic anemia in HIV patients with G6-PD deficiency. The current modeling uses the simplifying assumption that there is a single level of relatively specific topics. In contrast to the ephemeral lifespan of authors, papers, once written, exist forever.

The set of authors is interlinked via undirected coauthorship relations. Papers are interconnected via directed "provides input to" links. Authors and papers are interlinked via directed "consumed" links denoting the flow of information from papers to authors as well as directed "produced" links representing the act of paper generation by authors. Note that the decision to direct links according to the flow of information reverses the direction of the commonly used "cited by" link. The in-degree of a paper node refers to its number of references and the out-degree to its number of received citations.

Coauthor, citation, consumed, and produced links, once created, are permanent. Coauthorship links may become stronger as more and more papers are coauthored together. The number of provides input to links representing received citations may grow over time. Note that citation links can be created to any existing paper. However, coauthorship links can only be made to

[††]Simkin, M.V. & Roychowdhury, V.P. (2003) *Condens. Matter*, cond-mat/0305150 (abstr.).

[‡‡]Vazquez, A. (2001) *Condens. Matter*, cond-mat/0006132 (abstr.).

```
// Initialization
generate #_papers papers and assign a random topic to each paper;
generate #_authors authors and assign a random topic to each author;
randomly assign #_co-authors+1 authors to papers of the same topic;
// Simulation
for each year do {
   add #_new_authors new authors, deactivate authors older than #_author_age;
   for each topic do {
      randomly partition set of authors into author_groups of size #_co-authors+1;
      for each author_group do {
         for each new_paper to be produced, do {
            generate new_paper;
            randomly select #_read_ papers from existing papers;
            get all references of read_ papers up to #_reference_path_length;
            for each new_paper_reference do {
               select a time_slice from (start year to curr_year-1) with probability given in aging_function;
               randomly select a paper published or cited in this time_slice; as a new_paper_reference;
               add the new_paper_reference to new_paper;
            }
         }
      }
   }
   add all new papers to the set of existing papers;
   add new links to author and paper information;
}
```

Fig. 1. Process model in pseudo code. If no topics are considered then the number of topics is one, i.e., all papers and authors have the same topic. If no coauthors are considered then each paper has exactly one author. If the reference path length is 0 then no references are considered for citation. If no aging function is given then all papers have the same probability of getting selected.

currently active authors. For simplicity, each paper has a fixed number of authors and a fixed number of references.

Topics are randomly assigned to authors at initialization time. Papers inherit the topic of their author(s). In the current instantiation of the model, each author and each paper has exactly one topic. The initial number of used topics is typically lower than the total number of available topics. Eventually all of the topics are covered as new authors with randomly assigned topics are added gradually. Consumed-produced relationships among papers and authors are restricted to authors and papers within the same topic. Although this is a rather unrealistic assumption, it parsimoniously models the fact that authors from one knowledge domain typically do not frequently read journals, attend conferences, coauthor, or interact from/with other domains.

In the general model, the number of papers produced by an author would be a random variable. However, for the current instantiation of the model, a single fixed number of papers per author per year is assumed. We are aware that this is unrealistic and will not result in a Gaussian or power law distribution for the number of coauthors nor the number of papers published per author. Hence, only the properties of the paper citation network will be validated against the PNAS data set.

During model initialization, a set of authors and a set of papers with randomly assigned topics are generated (see pseudocode in Fig. 1). Subsequently, a predefined number of coauthors sharing the same topic is randomly selected and assigned to each paper via produced by links. All papers have authors but there may be authors that have not yet produced papers. There are initially no coauthor or paper citation links, making it advantageous to start the model at least 1 year earlier than the time period of interest.

At each time step (year), a specified number of new authors A' is created with a specific time stamp and added to the set of existing authors $A_t = A' \cup A_{t-1}$. A number of authors can be deactivated as well and subtracted from the set of authors. Subsequently, each author in set A_t randomly identifies a set of coauthors, reads a specified number of randomly selected papers from within his/her topic, and produces a specified number of new papers. Each new paper will cite a fixed number of existing papers, a number frequently stipulated by publishers and constrained by the number of pages available per journal. To select the papers cited, authors consume (read) a rather small set of papers because of finite cognitive and time constraints. To model the local networking activity of authors, the number of levels of paper references that are followed up by an author is modeled explicitly. If it is zero, then only the papers that an author reads in a year, the set P_0, can be cited. If it is two, then an author can cite any paper that they have read, any paper P_1 that is cited in one of the P_0 papers, or any paper that is cited in any of the papers in set P_1. With each deeper level of references, the set of papers, P', is added to $P_r = P' \cup P_{r-1}$. The set of papers available as citation references for a given year t and depth of reference level r is denoted by $P_{r,t}$.

A parameterized model was implemented in Java to simulate the simultaneous growth of interlinked coauthor networks and paper citation networks as described above. The model takes as input parameters that specify the number of authors and papers created in the initial year as well as the number of topics, the number of authors to be deleted per year, the number of papers an author produces per year, the number of papers cited by a new paper, the number of coauthors, the number of levels references are followed up, and the parameters of the aging function.

The model can be started with or without topics, coauthorships, following up of references, aging of papers, or any combination of these variables. A small author–paper network for topics only and for coauthors only is shown in Fig. 2 *a* and *b*, respectively. Fig. 2*a* shows three unconnected topic-based author and paper networks. During the simulation, authors read, cite, and produce papers from their topic area exclusively. Given that authors exclusively coauthor with authors within their own topic, each paper has exactly one topic. Author a4 was assigned a topic for which two papers existed after the initialization and hence both papers are assigned to the author. Later on the author generates one paper each year that cites the author's own work on the topic. Without any new authors or new topic areas for existing authors, all subsequently produced papers will belong to one of these three topic areas. With authors reading papers only of their own topic area, there will never be any links

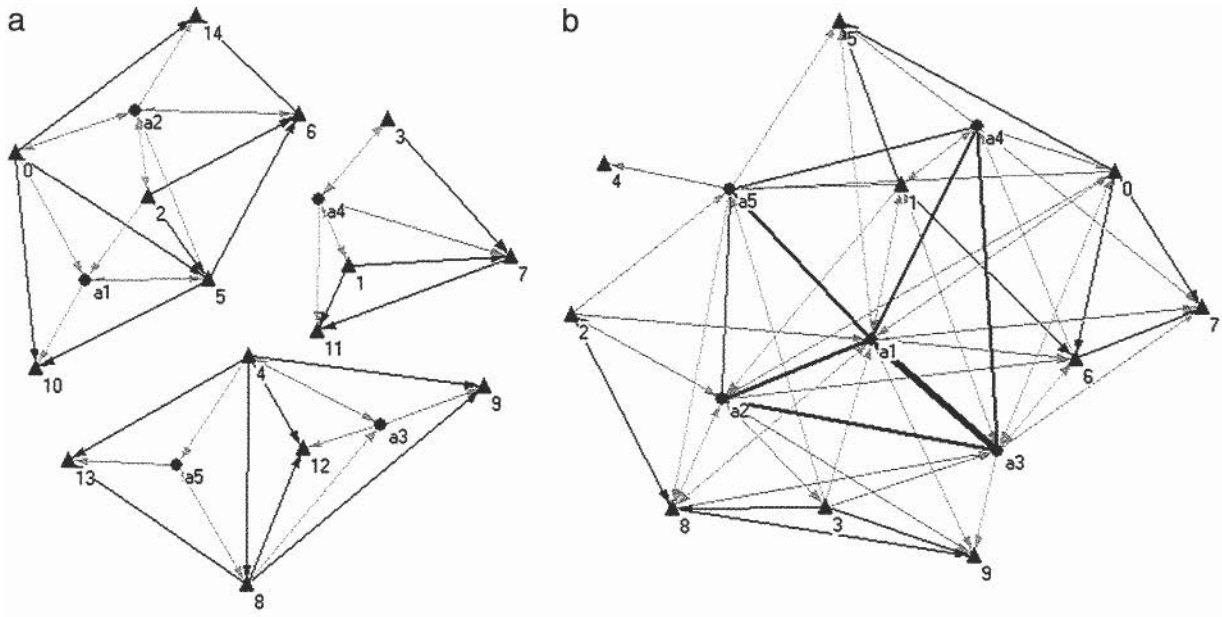


Fig. 2. Author–paper network generated by using the model with topics only (a) and coauthors only (*b*). The model was started with five topics, authors, and papers and run for 2 years. In each year, each author produces one paper, which cites two earlier papers. No authors were added or deactivated. The resulting networks has five authors (labeled a1–a5, blue circles) and 15 or 9 papers (labeled 0, 2, 3 . . . , red triangles). Papers are linked via red directed provided input to links. Authors are connected by blue coauthorships links. Light green indicates directed links denoting the flow of information from papers to authors and from authors to new papers via consumed and produced relations.

between the three topical clusters. If authors do not coauthor then there are no coauthor links.

If coauthorship is simulated, then each paper is authored by a predefined number of authors. At each time step, each author will select a number of coauthors to produce papers, and each produced paper has multiple authors. The model stops when the number of specified time steps is reached. In the network shown in Fig. 2*b* each newly generated paper has exactly two authors. Blue, undirected lines represent coauthorships. Line thickness indicates the number of papers that have been coauthored together, e.g., a1 and a3 coauthored several times. The total number of papers produced each year is lower than in Fig. 2*a* because two authors produce one paper together. If a topic area has fewer authors than needed for the collaborative production of a paper then no papers are produced.

If no references and no aging are considered then references are randomly selected from the set of papers that a coauthor team selected for reading. When references in papers are followed up then authors consider not only the papers they read as potential reference candidates but also papers linked to those via citation references up to a path of a certain length. Thus, a paper that was cited five times has six chances (or tokens) to get selected. The resulting paper citation network has some nodes, typically older papers, which are very highly cited, whereas the majority of papers are rarely, if at all, cited; see Fig. 3*a*.

If references as well as aging are considered, then the probability of paper y being cited, $P(y)$, corresponds to the normalized sum of the aging-dependent probability for each of its tokens,

$$P(y) = \frac{\sum_{t=1}^{n} \sum_{i \in P_{r,t} \wedge i=y} W(t)}{\sum_{t=1}^{n} \sum_{i \in P_{r,t}} W(t)},$$

where n is the total number of years considered. Hence a paper that was published in year y and received four citations in year $y + 1$ and two citations in year $y + 2$ has seven tokens that are weighted by the probability value for each year. The probability of citing a paper written t years ago can be fit by a Weibull distribution of the form

$$W(t) = cab^{-a}t^{(a-1)}e^{-\left(\frac{t}{b}\right)^{a}},$$

where c is a scaling factor, a controls the variability of distribution, and b controls the rightward extension of the curve. As b increases, the probability of citing older papers increases. For the present purposes, a small value of b represents a strong aging bias that favors citing papers that have been published recently. For small values of b, the function predicts very few citations for older papers. The introduction of aging offsets the rich-get-richer effect that favors the citation of older papers that have already been frequently cited.

The parameters specified in the input script file provide flexibility to fit the model output to diverse data sets. The model is used to fit the PNAS data in the next section. Later, we examine the influence of aging, reference path length, and number of topics on the structure of interlinked coauthor and paper citation networks.

Model Validation

To validate the TARL model, a 20-year (1982–2001) data set of PNAS was used. Subsequently, we describe the data set, select a set of model parameters, and compare the model output with the PNAS data set in terms of network properties.

The PNAS Data Set. The PNAS data set contains 45,120 regular articles. The number of unique authors for those papers is 105,915. Table 1 provides counts of the total number of papers, authors, references, and citations received by all of the papers for each of the 20 years, as well as the average number of coauthors per author. The average number of papers published per author

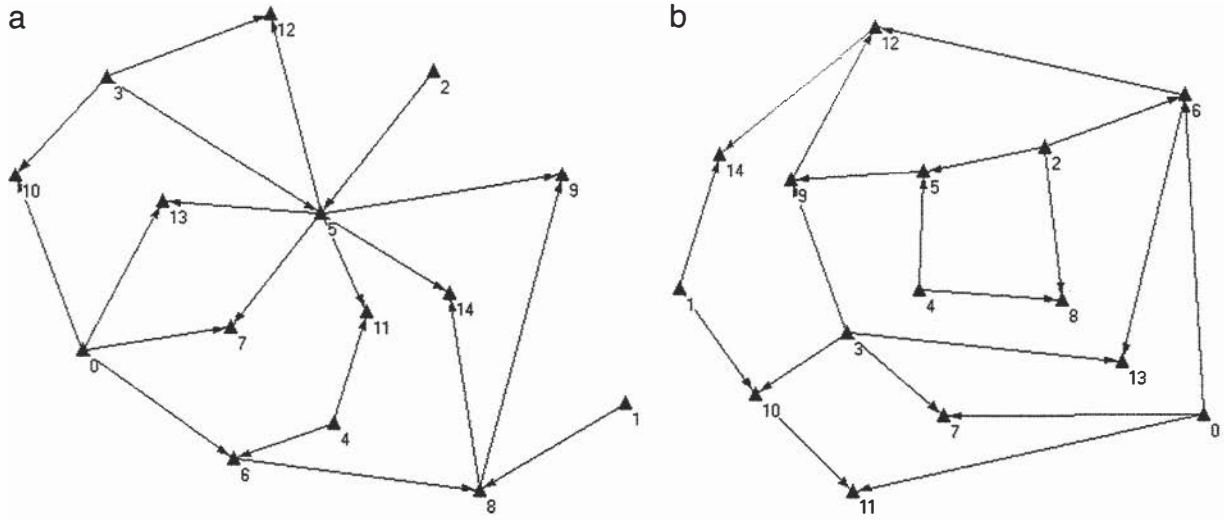


Fig. 3. Paper network without aging (a) and with aging (*b*). Without aging, older papers are more likely to provide input to younger papers; i.e., they are attracting most of the citation links. For example, in a, paper no. 0 generated at initialization time and paper no. 5 generated in year 1 provide input to four and six papers, respectively.

and the average number of references and citations per paper can be easily derived from Table 1. Note that the citation counts, particularly for younger papers, are artificially low because they have not existed in the literature long enough to garner many citations. Table 1 also provides information on the number of citations received from papers within the PNAS data set in the right-most column. Only those intra-PNAS citation links will be modeled. The paper most highly cited by papers within the set received 285 citations.

Fig. 4 visualizes the limited coverage of the data set. It neither contains all work by many authors for the 20-year time span as they may have published in other venues as well, nor does it provide information about citations received from PNAS papers published past 2001 or non-PNAS papers. References to papers outside the 20-year data set will be ignored.

Table 1. PNAS statistics in terms of total number of papers (#p), unique authors (#a), references (#r), citations received per paper (#c), number of coauthors per paper (a#ca), and the number of citations (#c_win) within the PNAS data set for each year

Year	#p	#a	#r	#c	a#ca	#c_win
1982	1,669	5,201	46,665	56,690	3.92	6,749
1983	1,611	5,142	46,685	161,437	3.98	7,188
1984	1,695	5,583	49,834	174,161	4.22	6,928
1985	1,846	6,325	55,662	191,750	4.38	7,425
1986	2,042	7,209	64,379	218,229	4.76	7,985
1987	1,924	7,061	59,110	207,729	4.88	7,340
1988	2,035	7,471	63,116	215,227	4.8	7,547
1989	2,088	7,959	65,883	215,437	5.01	7,386
1990	2,066	8,031	66,019	207,138	5.15	7,089
1991	2,382	9,559	77,740	223,102	5.25	7,511
1992	2,500	9,812	80,949	211,238	5.29	6,932
1993	2,413	9,770	79,848	193,867	5.55	5,979
1994	2,600	10,656	86,176	187,353	5.56	5,910
1995	2,476	10,429	82,021	151,249	5.66	4,922
1996	2,765	11,803	99,061	148,622	5.96	5,013
1997	2,618	11,255	96,788	122,908	6.12	4,290
1998	2,711	12,328	100,973	107,764	6.48	3,580
1999	2,603	12,182	97,018	76,080	6.69	2,453
2000	2,501	12,201	94,181	44,131	7.6	1,354
2001	2,575	13,038	97,450	16,357	8.4	422
Total	45,120		1,509,558	3,230,469		114,003

Table 2 lists small-world properties and power law exponents for diverse coauthor and paper citation networks. The values for the PNAS data set under examination and the simulated paper citation network are also given.

Note that for undirected coauthor networks, the in-degree of a node equals its out-degree and hence the exponents for both distributions are identical. For directed paper citation networks, the number of references is rather small and constant. As typical, only the in-degree distribution (received citations) are considered (7) and the γ reported values for paper citation networks characterize the in-degree distribution. For paper citation networks, we do not report the value for the characteristic path length as it reflects the time duration of the sample but little about the structure of the network.

Based on these values, the PNAS data set can be classified as a medium-sized data set that has a similar average node degree $\langle k \rangle$, path length l, cluster coefficient C, and power law exponent γ to the networks previously examined. The $\langle k \rangle$ value of the paper citation network is rather low. The total number of links within the PNAS citation network is 114,003. On average, each

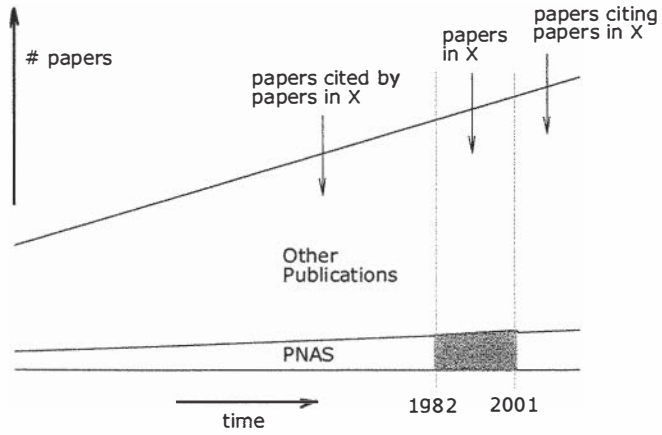


Fig. 4. Coverage of the PNAS data set in terms of time span, total papers, and complete authors' work.

Table 2. Properties of coauthor and paper citation networks comprising number of nodes n, average node degree $\langle k \rangle$, path length l, cluster coefficient C, and power law exponent γ

Network	n	$\langle k \rangle$	l	C	γ
Coauthorship networks					
LANL	52,909	9.7	5.9	0.43	—
MEDLINE	1,520,251	18.1	4.6	0.066	—
SPIRES	56,627	1.73	4.0	0.726	1.2
NCSTRL	11,994	3.59	9.7	0.496	—
Mathematics	70,975	3.9	9.5	0.59	2.5
Neuroscience	209,293	11.5	6	0.76	2.1
PNAS	105,915	8.97	5.89	0.399	2.54
Paper citation networks					
ISI	783,339	8.57	—	—	3
PhysRev	24,296	14.5	—	—	3
PNAS	45,120	3.53	—	0.081	2.29
SIM	37,114	2.13	—	0.074	2.05

Values for the first four coauthorship networks are taken from refs. 28–30. Math and neuroscience network were analyzed (21); Redner (31) reported the paper citation network values for ISI and PhysRev. Values for PNAS and simulated network data were acquired by the authors.

paper receives about three citations from another paper in this data set. The coauthor network has 472,552 links.

The power law exponent for the PNAS coauthor network is 2.54 and seems to match the values reported for other networks well. It accounts for 91% of the variance in the relation between number of coauthors and frequency of occurrence. The number of authors with very few coauthors is less than predicted by a power law relation, and the number of authors with a moderate number of coauthors is more than predicted. The best-fitting power law exponent for the paper citation network is 2.29, and the power law accounts for 87% of the variance. The systematic deviations from a power law are that most cited papers are cited less often than predicted by a power law, and the less cited papers are cited more often than predicted. A plausible account for these deviations are that networks in which aging occurs, e.g., actor networks, friendship networks, but also paper citation networks, show a connectivity distribution that has a power law regime followed by an exponential or Gaussian decay or have an exponential or Gaussian connectivity distribution (12). Newman (30) showed that connectivity distributions of coauthor networks from astrophysics, condensed matter, high energy, and computer science databases can be fitted by a power law form with an exponential cutoff. Following this lead, we fit a power law with exponential cutoff of the form $f(x) = Ax^{-B}e^{x/C}$. This function provided an excellent fit to the PNAS paper citation network with values of $A = 13{,}652$, $B = 0.49$, and $C = 4.21$ ($R^2 = 1.00$).

Model Initialization. The statistical properties of the PNAS data set were used to select the initialization values for the model. The model was run with topics, coauthors, references, and aging for 21 years covering 1981–2001. The year 1981 was used for initialization purposes. In 1981, 4,809 authors and 1,624 papers covering 1,000 topics were generated (see discussion of the linear relation between cluster coefficient and topics in the next section). In accordance with the PNAS data, the number of active authors was increased by 430 each year. Note that this increase in authors is caused mostly by external factors such as funding, which are not modeled in the current simulation and hence have to be supplied by hand. Even though 20 years is a rather large time span, the simplifying assumption was made that all authors remained alive/active. Although the number of coauthors increases continuously over time we decided to use the average value of 4. Hence the number of authors per paper is five. One paper is produced by each author per year. The average number of references per paper to papers within PNAS was set to 3 as determined by the actual data. One level of references was considered and the Weibull aging function was used, with a parameter value of $b = 3$, providing a 12-year time window in which papers are cited.

Statistics. Simulated data have been compared to the PNAS article data set in terms of total number of papers, unique authors, and citations received per paper for each year given in Table 1, as well as in terms of their small-world properties. Interestingly, the total number of papers in the simulation is slightly lower than the actual PNAS data. This is because authors who do not manage to find a sufficient number of coauthors in their topic area will not produce any paper in this particular year. Similarly, papers that are produced in a topic area with very few papers will not be able to reference the called for number of three papers. Hence the average degree $\langle k \rangle$ is slightly lower than the value observed for the PNAS paper network.

As shown in Fig. 5*a* the number of papers published in PNAS increases slowly but steadily over the 20-year time period. The number of authors publishing in PNAS increases more rapidly than the number of papers because the average number of coauthors per paper increases from 3.49 in 1982 to 5.42 in 2001. The simulation assumes a linear increase in authors over time,

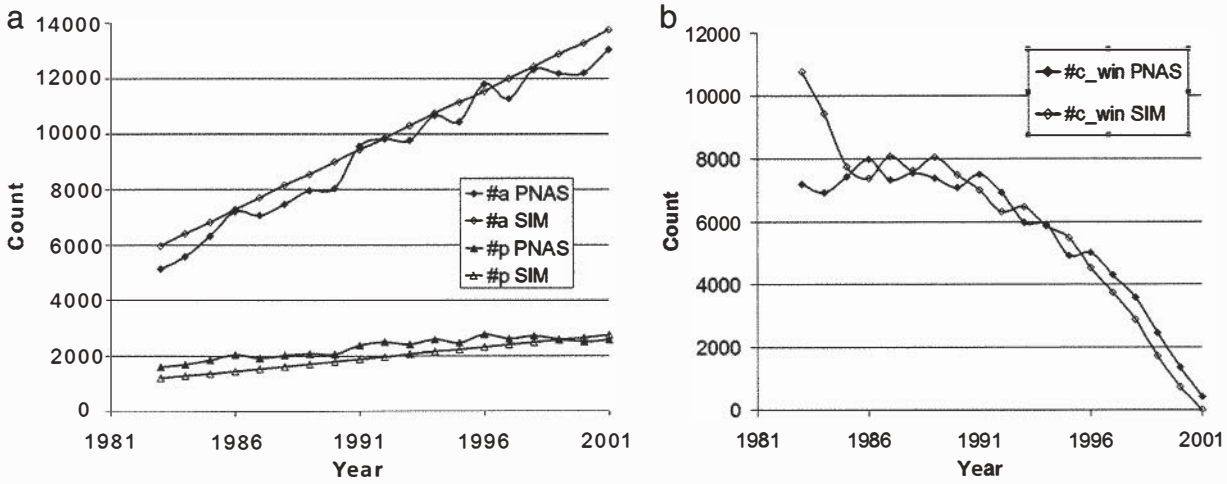


Fig. 5. Total number of actual and simulated papers (#p) and authors (#a) (a) and received citations (#c_win) (*b*).

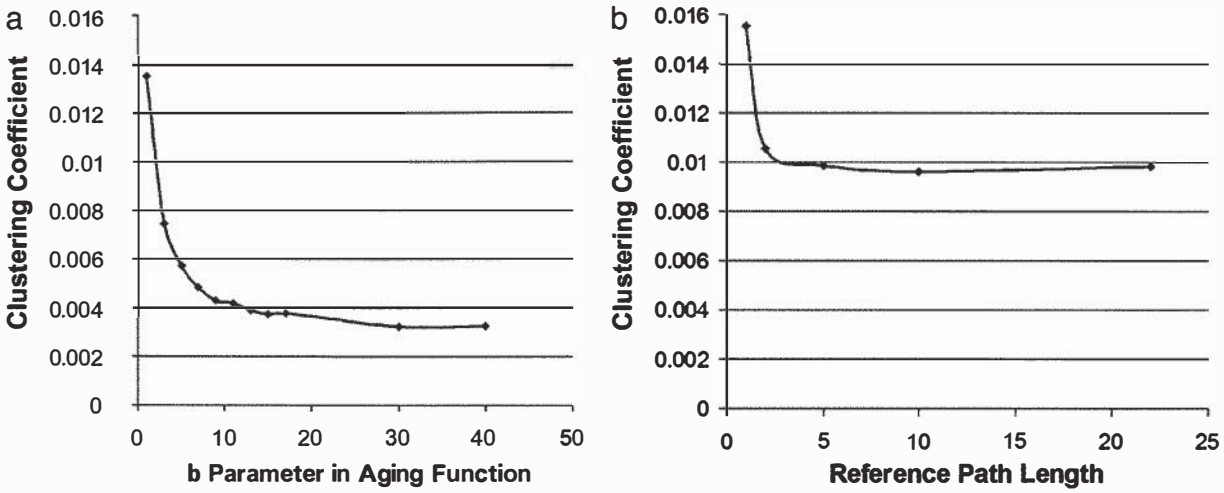


Fig. 6. Cluster coefficient as a function of the aging function (a) and the reference path length (*b*).

but the increase in the number of papers produced naturally comes out of this increase in authors.

The average number of received citations for each year is displayed in Fig. 5*b*. The model closely tracks the number of actual citations for all years after 1984. The fit for the first 2 years is poor because the model has no initial citation links nor record of papers before 1981. Given that no papers before 1981 are available as references, papers published in early years of the simulation receive a disproportionately large number of citations. This effect fades away in 1985 as the aging function selects mostly papers published in the last 12 years and papers published in the last 7 years have a particularly high probability of being cited. The total number of citations received by papers within the PNAS data set is 111,341. The artifacts during the initial phase of the model run could be eliminated by starting the model 10 years earlier and analyzing only the final 20 years. However, we believe the graph in Fig. 5*b* nicely illustrates the influence of aging and the model in action. Both actual and modeled data sets reflect the fact that younger papers have shorter periods of time in which to draw citations.

Network Properties. This section discusses the fit of the simulated paper citation network to the PNAS data in terms of small-world properties as well as the power law exponent γ for the paper connectivity graph. Results for the best-fitting model parameters are reported in the last row of Table 2.

The cluster coefficient for simulations in which all authors and papers have the same topic is rather low. The cluster coefficient increases considerably if topics are considered; see also the discussion on topics in the next section.

The simulation with 1,000 topics and an aging parameter of $b = 3$ provides a good fit to the PNAS data set in terms of the distribution of citations. The model data R^2 was 0.996, which is substantially better than the best-fitting power law to the PNAS data ($R^2 = 0.87$), and almost as good as the best-fitting power law with exponential tail ($R^2 = 1.00$). As with the PNAS data, the simulated data were fit much better with a power law with exponential tail ($R^2 = 0.999$) than simple power law (0.987). Although the simulation does not fit the PNAS data any better than the power law with exponential tail, it does provide a process model for why this functional relation applies. Very highly cited papers are more rare in the PNAS and simulated data sets than predicted by a power law because of the bias toward citing recent papers. The tendency for highly cited older papers to attract still more citations is offset by a counteracting tendency to cite recent papers.

The Influence of Model Parameters

This section discusses the influence of different TARL model parameters on network properties such as cluster coefficient and power law exponent for the citation distribution.

Interestingly, the number of topics is linearly correlated with the clustering coefficient of the resulting network: $C = 0.000073$ * no. topics. Hence our knowledge about the clustering coefficient in the PNAS network governed the choice of 1,000 topics. The linear relation entails a desirable property of the simulation; a simple method exists for creating networks with a specific degree of clustering.

Topics also influence the power law exponent for the citation distribution. Increasing the number of topics increases the power law exponent as authors are now restricted to cite papers in their own topics area. By dividing science into separate fields, the global rich-get-richer effect is broken down into many local rich-get-richer effects, leading to a more egalitarian distribution of received citations.

Aging refers to the distribution of probabilities for papers being cited by new papers. The influence of the b value used to generate different Weibull aging functions is shown in Fig. 6*a*. The aging distribution observed in the PNAS data was used to determine the parameter value $b = 3$, marked with a star in Fig. 6. By increasing b, and hence increasing the number of older papers cited as references, the clustering coefficient decreases. This effect suggests a second kind of clustering that parallels the strong topic-induced clustering described previously. Papers are not only clustered by topic, but also in time, and as a community becomes increasingly nearsighted in terms of their citation practices, the degree of temporal clustering increases.

Last but not least, the length of the chain of paper citation links that is followed to select references for a new paper also influences the clustering coefficient. The dependence of the clustering coefficient on the reference path length is given in Fig. 6*b*. This result indicates that temporal clustering is ameliorated by the practice of citing (and hopefully reading!) the papers that were the earlier inspirations for read papers.

Note that the aging and reference path length examinations were conducted for 200 topics.

Discussion and Outlook

This article presented results on modeling the simultaneous evolution and structure of author–paper networks. Although prior research has described the associations among different

scientific structure (e.g., authors, publications, topics, web) (23), to our knowledge, nobody has yet attempted to model the simultaneous growth and dynamic interactions of multiple networks dealing with scientific output.

Models based on preferential attachment assume that new papers are linked to highly connected (cited) papers and new authors tend to coauthor with already highly interlinked authors. However, in today's dynamic scientific world of increasing specialization, an overview of the connectivity of a scientist or paper is not available to authors (even experts in a field). Instead, each author can be seen as a part of a complex network with local connections. Each author interacts directly only with a rather limited number of other authors and papers. However, papers that are cited frequently have a higher probability of being cited again. Similarly, authors that are highly interconnected with other authors in social networks are likely to attract more coauthors if we assume that authors tend to coauthor with coauthors of their coauthors. The presented model uses the reading and citing of paper references as a grounded mechanism to generate paper citation networks that are approximately scale-free. Moreover, the particular deviations from scale-free properties are well predicted by a version of the model that incorporates a bias to cite recent papers and a scientific community that is subdivided into specialized topics. The model parameters that governed these two factors were b that reflects that influence of aging and "number of topics" reflecting the degree of splintering within science. The values for these parameters were not freely fit to the citation distribution data. Instead, the number of topics was selected to approximate PNAS's clustering coefficient, and b was selected to provide the optimal Weibull fit to PNAS's distribution of citations as a function of lag in years. Thus, the highly respectable model-data fit involving the number of citations and their frequency is impressive because it involves no true free parameters.

The incorporation of topics and recency bias was instrumental in achieving the qualitative violations of a power law distribution. There are fewer papers that receive a large number of citations than is predicted by a power law, because the bias toward citing recent papers offsets the rich-get-richer effect that generates a power law relation. It is difficult for a well cited paper to continue to receive additional citations as it ages. The citation bias toward recent papers combines with the within-topic citation constraint to create citations networks that have high degrees of clustering by both topic area and time. These model assumptions account for the observed citation distribution and suggest an interesting interplay between citation practices that lead to egalitarian versus lopsided distributions of citations.

For the sake of simplicity the number of papers produced by each author per year was fixed and a fixed number of coauthors were randomly assigned to each author. If coauthors preferably collaborate with coauthors of their coauthors, this would provide a grounded mechanism for the generation of small-world and approximately scale-free network structures analogous to their construction in the paper network. Similarly to the aging of papers, the "deactivation of authors" could also be modeled. If authors are more likely to cite papers of active authors, then the deactivation of all authors of a paper would decrease the "attraction" or "fitness" of a paper to receive citations by another paper. The deactivation of authors would also cause previous coauthors to search for new coauthors. Having authors coauthor across topics would lead to a more realistic interconnection of papers from different topic areas via citation links.

The productivity of an author may depend not only on his/her position in the author–paper network but also on available research funds, facilities, and students.

To give an example, consider the feedback cycle of authors, papers, and funding. Authors that manage to produce many high-quality papers also increase their chances of receiving funding. Funding in return enables authors to hire (better) graduate students or postdocs, which in turn increases the number of coauthors and the amount and quality of paper output and hence the likelihood of attracting still more funding.

This work greatly benefited from discussions with and comments from Kevin Boyack, Albert-László Barabási, Mark Newman, Olaf Sporns, Filippo Menczer, and the anonymous reviewers. Mark Newman made code available to determine the small-world properties of networks. Nidhi Sobti was involved in the analysis of the influence of model parameter values. Batagelj and Mrvar's PAJEK program was used to generate the network layouts (24). This work is supported by National Science Foundation CAREER Grant IIS-0238261 (to K.B.) and National Science Foundation Grant 0125287 (to R.L.G.).

1. White, H. D. & McCain, K. W. (1989) in *Annual Review on Information Science and Technology*, ed. Williams, M. E. (Elsevier, Amsterdam), Vol. 24, pp. 119–186.
2. van Raan, A. F. J. (1997) *Scientometrics* **38,** 205–218.
3. Börner, K., Chen, C. & Boyack, K. (2003) in *Annual Review of Information Science and Technology*, ed. Cronin, B. (Information Today/Am. Soc. Information Science and Technology, Medford, NJ), Vol. 37, pp. 179–255.
4. Albert, R. & Barabási, A.-L. (2002) *Rev. Mod. Phys.* **74,** 47–97.
5. Watts, D. J. (1999) *Small Worlds: The Dynamics of Networks Between Order and Randomness* (Princeton Univ. Press, Princeton, NJ).
6. Barabási, A.-L., Albert, R. & Jeong, H. (2000) *Physica* **281,** 69–77.
7. Dorogovtsev, S. N. & Mendes, J. F. F. (2002) *Adv. Phys.* **51,** 1079–1187.
8. Merton, R. (1973) *The Sociology of Science* (University of Chicago Press, Chicago).
9. Price, D. J. D. S. (1976) *J. Am. Soc. Information Sci.* **27,** 292–306.
10. Barabási, A.-L. & Albert, R. (1999) *Science* **286,** 509–512.
11. van Raan, A. F. J. (2000) *Scientometrics* **47,** 347–362.
12. Amaral, L. A. N., Scala, A., Barthelemy, M. & Stanley, H. E. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 11149–11152.
13. Garfield, E., Sher, I. H. & Torpie, R. J. (1964) *The Use of Citation Data in Writing the History of Science* (Institute for Scientific Information, Philadelphia).
14. Garfield, E. & Small, H. (1989) in *Innovation at the Crossroads Between Science and Technology*, eds., Kranzberg, M., Elkana, Y. & Tadnor, Z. (S. Neaman Press, Haifa, Israel).
15. Adams, J. & Griliches, Z. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 12664–12670.
16. Morris, S. A., Yen, G., Wu, Z. & Asnake, B. (2003) *J. Am. Soc. Information Sci. Technol.* **54,** 413–422.
17. Willinger, W., Govindan, R., Jamin, S., Paxson, V. & Shenker, S. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 2573–2580.
18. Erdos, P. & Renyi, A. (1960) *Hungarian Acad. Sci.* **5,** 17–61.
19. Albert, R., Jeong, H. & Barabási, A.-L. (2000) *Nature* **406,** 378–382.
20. Watts, D. J. & Strogatz, S. (1998) *Nature* **393,** 440–442.
21. Barabási, A.-L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A. & Vicsek, T. (2002) *Physica A* **311,** 590–614.
22. Kleinberg, J., Kumar, S. R., Raghavan, P., Rajagopalan, S. & Tomkins, A. (1999) in *Proceedings of the International Conference on Combinatorics and Computing:* Lecture Notes in Computer Science, eds. Imai, H., Lee, D. T., Nakano, S.-I., Tokuyama, T., Asano, T. (Springer, Berlin), Vol. 1627, pp. 26–28.
23. Menczer, F. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5261–5265.
24. Batagelj, V. & Mrvar, A. (1998) *Connections* **21,** 47–57.
25. Banks, D. L. & Carley, K. M. (1996) *J. Math. Soc.* **21,** 173–196.
26. Snijders, T. A. B. (2001) in *Sociological Methodology*, eds. Sobel, M. E. & Becker, M. P. (Blackwell, Boston), pp. 361–395.
27. Gilbert, N. (1997) *Sociological Research Online* Vol. 2 (No. 2). Available at www.socresonline.org.uk. Accessed Nov. 11, 2003.
28. Newman, M. E. J. (2001) *Phys. Rev. E* **64,** 016131.
29. Newman, M. E. J. (2001) *Phys. Rev. E* **64,** 016132.
30. Newman, M. E. J. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 404–409.
31. Redner, S. (1998) *Eur. Phys. J. B* **4,** 131–134.

The world of geography: Visualizing a knowledge domain with cartographic means

André Skupin*

Department of Geography, University of New Orleans, New Orleans, LA 70148

From an informed critique of existing methods to the development of original tools, cartographic engagement can provide a unique perspective on knowledge domain visualization. Along with a discussion of some principles underlying a cartographically informed visualization methodology, results of experiments involving several thousand conference abstracts will be sketched and their plausibility reflected on.

The question "Hasn't everything been mapped already?" is commonly posed to someone who calls himself a cartographer in the early 21st century. It would then typically be countered with reference to the ever-changing nature of what geographers like to call the "infinitely complex geographic reality," requiring vigilance in keeping ever-more-detailed geographic databases up-to-date. Where ever-growing geospatial data repositories, advanced computing power, and cognitive insights meet, cartographers are advancing scientifically in a field known as geographic visualization.

At the fringes of this activity, some cartographers have begun to attempt a combination of centuries of accumulated cartographic knowledge with modern computational approaches and cognitive insights, toward the visualization of nongeographic information. Examples for such nongeoreferenced data are the text document corpi held in digital libraries, user interaction logs created by Web applications, or biological data associated with genome mapping. In all of these cases, researchers of the interdisciplinary effort known as information visualization are engaged in the endeavor of making high-dimensional structures more directly accessible to the human cognitive system (1). Arguably, lessons from traditional cartography and transformation techniques derived from geographic information science would be applicable to many aspects of information visualization (2). This holds especially true in the context of 2D representations on screen or paper and in the even more narrowly defined, yet extremely popular, group of map-like information visualizations (3). Some results of this ongoing cartographic involvement are discussed here.

Implementation of Map-Like Knowledge Domain Visualization

A spatialization of the geographic knowledge domain is presented here on the basis of an analysis of abstracts submitted to the annual meeting of the Association of American Geographers. With all of the branches of geography participating at this meeting, the data set and resulting visualizations provide a fairly comprehensive snapshot of the geographic discipline, from established, well-publicized research fields to those only recently emerging. The goal is to implement a multilevel visualization, in which major research areas as well as finer nuances of geographic activity would be shown. There is a range of possible uses for such visualizations. Beginning geography students could be introduced to the topical structure of the discipline. Geographic researchers could see their own work in the context of broader disciplinary trends. Visualizations like this could ease collaboration in interdisciplinary research settings, and so forth.

The input data set consisted of 2,220 abstracts, as submitted by conference participants, stored in ADOBE pdf format on the conference compact disk. After conversion to a plain text format, each abstract's content was parsed into such components as title, author information, full text, and author-chosen keywords. Then, the text of abstracts was indexed against the full set of author-chosen keywords of all abstracts (4).

In the absence of citation information, the visualization methodology chosen in this project follows a straightforward content-based path (as opposed to exploiting an explicit citation link structure) based on vector-space modeling and use of the self-organizing map (SOM) method (Fig. 1). The methodology is thus related to a number of projects using a similar approach (5–7). However, there are also significant differences that, in combination, lead to visualizations bearing a distinctly cartographic mark (Fig. 2).

Following a standard vector-space implementation for the document corpus, a SOM consisting of a relatively large number of neurons is trained (4,800 neurons) so that unique 2D locations for individual documents can be derived (2,220 documents). Then, a hierarchical cluster solution involving all n-dimensional neuron vectors ($n = 741$) is computed to support a multiscale zoomable visualization (4).

Because natural language is the primary means by which scientific knowledge is formally disseminated and conveyed in many domains, meaningful labeling of geometric features ought to be not an afterthought but an integral part of knowledge domain visualizations. Contrary to common performance-oriented level-of-detail approaches, the aim here is to convey semantic aspects of the geographic domain in accordance with scale-dependent notions of global vs. regional vs. local structures. For example, the distinction of human geography and physical geography is a global one, whereas urban and industrial geography are regional flavors of human geography, and abstracts dealing with car manufacturing locations across the globe would form local structures. To this end, the extraction of scale-dependent label terms is particularly stressed. Determination of cluster labels is based on a weighting formula that extends the popular term frequency × inverse document frequency mechanism from its traditional use for individual documents (8) toward groups of documents. For a given cluster, this formula will tend to emphasize terms that appear often within and rarely outside of that cluster, accommodating very well the needs of a multiscale representation. When dealing with a small number of clusters (i.e., at a global scale), the derived label terms will be quite general, e.g., "climate" or "urban." For a large number of

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

Abbreviation: SOM, self-organizing map.

*E-mail: askupin@uno.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0307654100

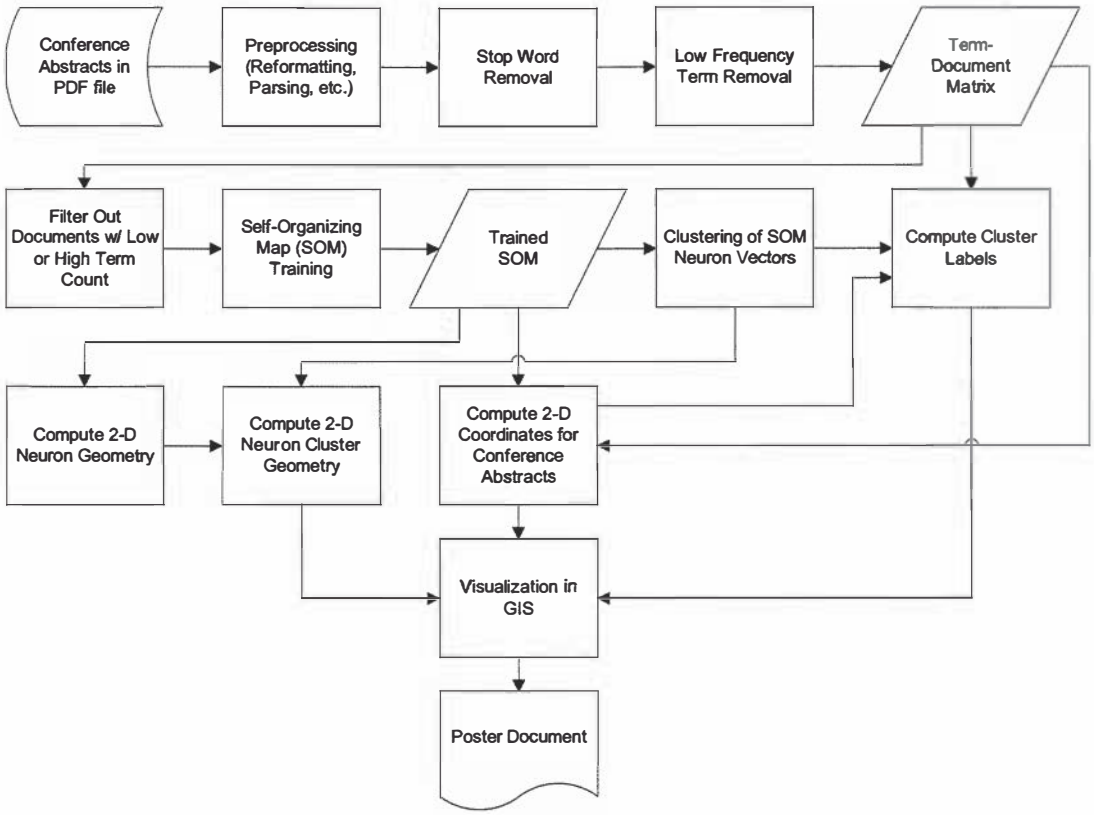


Fig. 1. Creation of a map-like visualization of conference abstracts using a self-organizing map and geographic information systems.

clusters (i.e., at a local scale), labels will correspond to much more specific areas of investigation in the geographic knowledge domain, e.g., "snowfall" or "redevelopment."

Cartography is essentially a science dealing with the transformation of spatial information (9). Following this paradigm, a number of geometric and topological transformations are applied to the raw geometric configuration produced by neural network training and, finally, symbolization occurs in off-the-shelf geographic information systems (GIS) software. This final step is driven by traditional cartographic considerations regarding visual hierarchies, here conveyed through color choices and manipulation of labels and line sizes. Label placement is performed automatically by GIS software.

Large-Format Knowledge Domain Visualization

With a display area of almost 12 square feet, the physical size of this visualization is more in tune with traditional cartographic output than snapshots presented in a journal paper (4) or

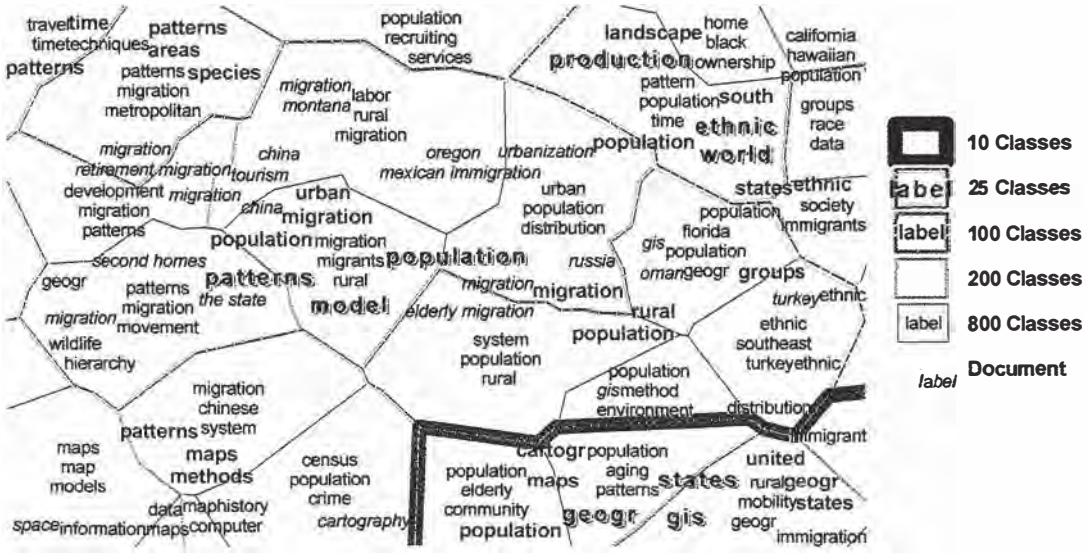


Fig. 2. Portion of a visualization of several thousand conference abstracts with simultaneous display of five cluster levels and individual documents.

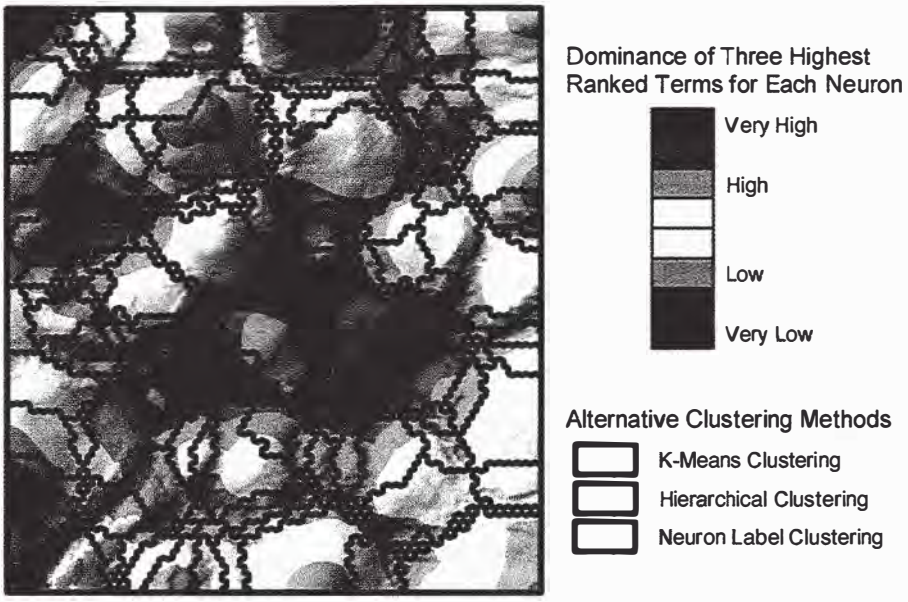


Fig. 3. Use of a term dominance surface to visually evaluate different clustering solutions.

interfaces heavily influenced by a limited screen size. It becomes possible to present multiple cluster levels simultaneously, making the use of hierarchical clustering particularly advantageous from a graphical point of view, because high-level cluster boundaries always also form lower-level boundaries. Rich labeling complements the extensive geometric structures created through this spatialization of conference abstracts, endowing the result with a remarkably map-like look (Fig. 2). A complete poster-size presentation of the result is available as Fig. 6, which is published as supporting information on the PNAS web site.

Creation of such large-format visualizations of knowledge domains is useful in various circumstances, especially in light of recent trends toward collaborative visualization (10). These efforts are complemented by a growing number of technologies that support the display of large-format visualizations, e.g., ImmersaDesk and DisplayWall. Interestingly, the major metaphors underlying the use of those technologies for visualization purposes, like drafting table or wallboard, correspond to traditional environments for cartographic map use.

Large displays on a static medium should not be easily dismissed either, especially when it comes to introducing novices to a knowledge domain and for establishing common ground among collaborating researchers. In those settings, these visualizations should be called "stable" rather than "static." This has been one of the enduring qualities of large-format geographic maps. For example, when introducing proposed changes to a land-use ordinance in a town hall meeting, large-size maps are not merely used for illustration. Their purpose is also not to simply transmit an encoded geographic "message" and certainly not to gain insight into a phenomenon, as is the case for most scientific visualizations. Instead, these maps help to establish a shared frame of reference, without which human-to-human communication would be much more difficult.

Much work remains to be done to uncover the relative cognitive value of large-format visualizations in general, including those depicting knowledge domains. Similarly, it remains to be tested whether and under which circumstances static depictions are indeed inferior to highly interactive systems, as seems to be presumed by most knowledge domain visualizations.

Clustering Methods

In considering the use of clustering methods, it should first and foremost be pointed out that the purpose of clustering in this line of research is not to discover optimal feature space partitions. Instead, clustering serves as a stepping-stone in the support of visual exploration toward domain comprehension. Note that visual exploration does not necessarily imply interactivity in a human–computer interaction sense. Arguably, viewers of richly symbolized but static knowledge domain visualizations are engaged in a process of visual exploration as well.

The choice of hierarchical clustering to create the large-format visualization discussed earlier is driven by the advantages it offers graphically, conceptually, and computationally. Its nested structure makes simultaneous display of multiple cluster levels feasible (Fig. 2). At lower cluster levels, only truly new geometric elements have to be added, as long as the cluster hierarchy is properly conveyed through a visual hierarchy (e.g., use of line thickness to convey cluster level). However, certain problems associated with hierarchical clustering are also apparent. The nested structure comes at the cost of a suboptimal tessellation of the n-dimensional input space. For example, notice the appearance of similar labels near the peripheries of neighboring clusters (Fig. 2), indicative of the tension between a strict partitioning mechanism and the continuous nature of the self-organizing map.

The SOM method, with its field-like continuous conceptualization of a high-dimensional information space, makes exact partitioning indeed difficult, especially in transitional zones. It would be useful to know how well different clustering methods perform under these conditions. Apart from standard statistical approaches, e.g., an investigation of within- and between-cluster variances, it is possible to use spatialization to that end as well.

Visual and computational overlays of various thematic layers on the basis of a common coordinate system have been a mainstay of geographic information systems philosophy for over three decades, since such operations were first proposed in a precomputer setting (11). Similarly, one could overlay different clustering solutions onto the same neuron geometry, which is illustrated by Fig. 3 for three cluster solutions:

(*i*) a k-means solution ($k = 25$);

(*ii*) one level derived from a hierarchical clustering tree (at the 25-cluster level);

(*iii*) a solution based on a method we call neuron label clustering, in which neighboring neurons are merged into clus-

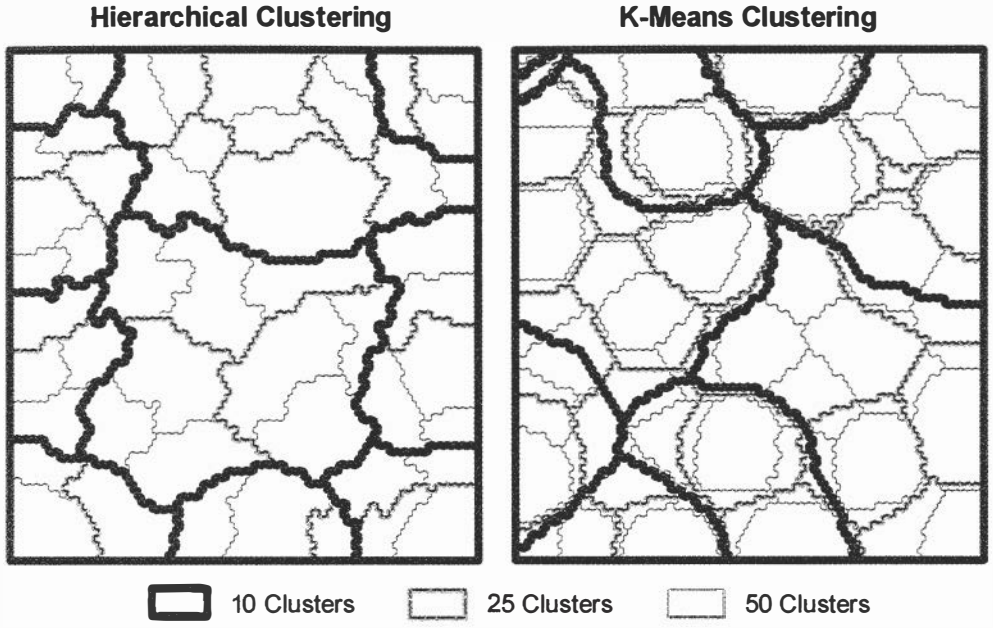


Fig. 4. Comparison of simultaneous display of multiple cluster levels based on two different clustering methods.

ters if their highest-weighted label terms are identical. This is similar to the clustering method proposed by Chen *et al.* (12).

Underneath, structures in the continuous information space are shown by means of a term dominance landscape, which expresses how dominant each neuron's top three label terms are with respect to all of the terms associated with a neuron. Because the training of SOM neurons is based on a dissimilarity/distance coefficient (in this case, the Euclidean measure), neighboring neurons will tend to have similar n-dimensional vectors associated with them, leading to a formation of extended mountain ranges. Higher "elevations," shown in brown tones, indicate a more coherently organized theme. Local minima may indicate a lack of distinct topical focus. "Clusters" incorporating those minima should thus be treated with some caution. Although superficially similar to other landscape-type knowledge domain visualizations, there are significant differences. Mountain ranges are formed by dominant combinations of keywords, i.e., major topics, across a large number of documents, which contrasts with a representation sometimes encountered of a majority of documents as local maxima (i.e., peaks) that seems to conflict with the continuous nature of the landscape metaphor. Formation of mountains is also not based on the density or number of documents that fall within its reach (13).

Valleys in the term dominance landscape correspond to transitional or overlapping topics between the dominant themes. This is again different from other landscape-type knowledge domain visualizations, where valleys mostly remain unpopulated by documents and must therefore be presumed to be void of meaning (13, 14).

Each clustering approach has distinct characteristics. Although the nested structure of hierarchical clustering has obvious advantages for graphic design and interaction, it has a tendency to cut through landscape features without obvious justification. The k-means method merges neighboring neurons into relatively evenly shaped and sized chunks, related to its use of the same objective function as the standard SOM training algorithm used here.

Of the three methods, the neuron label clustering approach matches the dominance landscape best, which makes sense because weighted term labels form the basis for computation of those two layers. Note how closely cluster boundaries follow "valley" features in the landscape, whereas "mountains" are enclosed. However, it offers the least control over granularity, which makes it difficult to create multiscale interfaces for exploration of knowledge domains.

Contrary to this, cluster levels in hierarchical and k-means clustering can be precisely chosen, as shown in Fig. 4. As mentioned earlier, the nested structure of hierarchical clustering reduces graphic and conceptual complexity (although we are not aware of human subject studies specifically investigating this issue). The k-means solutions appear graphically more complex, with plenty of overlapping clusters at different levels of k. On closer examination, some interesting observations emerge. Notice how some of the clusters at the 50-cluster level remain encircled and undivided by boundaries at the 25- and 10-cluster level, indicating agreement among different k-means solutions regarding these core areas. Interestingly, those cluster cores correspond to the major mountain ranges in the term dominance landscape (compare Fig. 3). On the other hand, peripheral clusters are formed at the 50-cluster level that are bisected at higher cluster levels. Those peripheral clusters correspond to either subtopics (i.e., divisions of larger topics), indicated by minor peaks within larger mountain ranges, or transitional/overlapping themes, shown as valleys in the term dominance landscape.

Compared to hierarchical clustering, the k-means method offers more optimal partitioning. On the other hand, it provides much better granularity control than neuron label clustering. Fig. 5 offers one suggestion for leveraging those characteristics while eliminating the complexity caused by cluster boundaries that do not coincide across multiple scale levels. The term dominance landscape is here combined with a labeled k-means solution, in which the cluster boundaries themselves are not shown explicitly but are at work in the background to automate placement of the computed cluster labels. Font size expresses the rank of a label term for that cluster. A semantic zoom operation is illustrated, during which a switch from a 25-cluster to a 100-cluster solution occurs. The mountain range labeled "population" is now shown in greater detail, breaking up related research topics into smaller categories, labeled "ethnic" to the right of the main peak and "migration" to its left. The location of these subcategories is significant, because the extensive use of

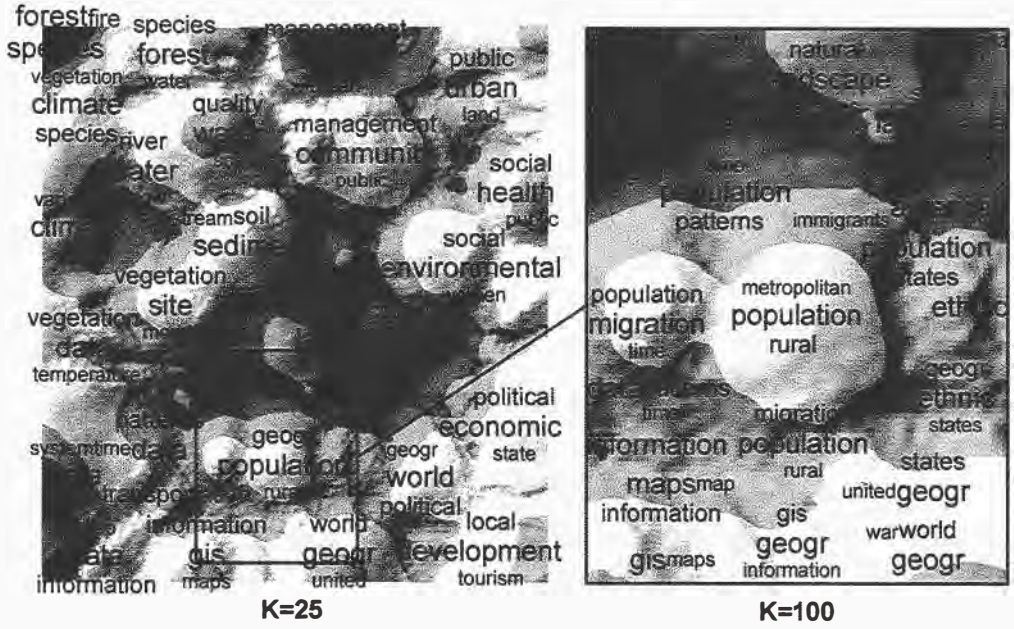


Fig. 5. Use of *k*-means clustering in combination with a term dominance landscape to support semantic zooming.

computational tools in migration studies warrants a position between the core population peak and the regions in the lower left of the global map focused on computational methodologies. This is quite different from studies of ethnic issues, which are typically grounded in a qualitative descriptive research paradigm, like many of the topics associated with the right half of this spatialization.

In summary, the purpose of clustering in knowledge domain visualization is not a provision of the "single best" and "true" partition of a domain, but rather one that may be useful under given circumstances. The examples discussed in this section demonstrate that the purpose of spatialization in the mapping of knowledge domains could extend beyond the creation of end-user tools. The computational procedures underlying multiscale visualizations may themselves be subject to visual inspection, and the resulting insights can inform the development of new or improved domain visualization methods.

Conclusion

This paper is largely driven by a desire to instigate reflection on the promise of the geographic metaphors and cartographic techniques that seem at the heart of so many knowledge domain visualizations. It raises important questions about the design of knowledge domain visualizations, such as: How far can we go in pursuit of cartographic metaphors? Is interactivity always necessary? Is there a role for static visualization in supporting discourse on the state and evolution of knowledge domains? Does the cognitive plausibility of certain visual approaches (e.g., a nested hierarchical structure) override a potential lack of computational plausibility? What would be the value of a convergence between knowledge domain visualizations and recent collaborative visualization developments?

This paper has demonstrated the possibility of creating large-format knowledge domain visualizations that emulate many aspects of traditional geographic depictions. Abstraction and scaling remain some of the most promising areas of cartographic influence on knowledge domain mapping efforts. In this context, this paper has presented an approach, informed by geographic information science, for the use of visual overlays to compare and validate different cluster techniques. The discussed techniques could of course be applied to similar data from other knowledge domains, as has been demonstrated elsewhere (15). We are currently developing a system aimed at providing a streamlined work flow for the creation of map-like knowledge domain visualizations. Future experiments involving both computational and human subject methodologies will help shed further light on the specific means for implementing useful map-like knowledge domain visualizations.

The research presented here is partially supported by the Louisiana Board of Regents Support Fund, Grant LEQSF(2002-05)-RD-A-34.

1. Card, S. K., Mackinlay, J. D. & Shneiderman, B. (1999) *Readings in Information Visualization: Using Vision to Think* (Morgan Kaufmann, San Francisco).
2. Skupin, A. (2000) in *InfoVis 2000* (Institute of Electrical and Electronic Engineers Computer Society, Salt Lake City), pp. 91–97.
3. Skupin, A. (2002) in *Visual Interfaces to Digital Libraries*, Lecture Notes in Computer Science, eds. Börner, K. & Chen, C. (Springer, Berlin), Vol. 2539, pp. 161–170.
4. Skupin, A. (2002) *IEEE Computer Graphics and Applications* **22,** 50–58.
5. Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, T., Paatero, V. & Saarela, A. (1999) in *Kohonen Maps*, eds. Oja, E. & Kaski, S. (Elsevier, Amsterdam), pp. 171–182.
6. Lin, X. (1992) in *IEEE Visualization '92* (Institute of Electrical and Electronic Engineers Computer Society Press, Los Alamitos, CA), pp. 274–281.
7. Rushall, D. & Illgen, M. (1996) in *InfoVis 1996* (Institute of Electrical and Electronic Engineers Computer Society Press, Los Alamitos, CA), pp. 100–107.
8. Salton, G. (1989) *Automated Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer* (Addison–Wesley, Reading, MA).
9. Tobler, W. (1979) *Am. Cartogr.* **6,** 101–106.
10. Brewer, I., MacEachren, A. M., Abdo, H., Gundrum, J. & Otto, G. (2000) in *InfoVis 2000* (Institute of Electrical and Electronic Engineers Computer Society, Salt Lake City), pp. 137–141.
11. McHarg, I. (1969) *Design with Nature* (Natural History Press, Garden City, NY).
12. Chen, H., Schuffels, C. & Orwig, R. (1996) *J. Visual Commun. Image Rep.* **7,** 88–102.
13. Boyack, K. W., Wylie, B. N. & Davidson, G. S. (2002) in *Visual Interfaces to Digital Libraries*, Lecture Notes in Computer Science, eds. Börner, K. & Chen, C. (Springer, Berlin), Vol. 2539, pp. 145–158.
14. Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A. & Crow, V. (1995) in *InfoVis 1995* (Institute of Electrical and Electronic Engineers Computer Society, Atlanta), pp. 51–58.
15. Börner, K., Chen, C. & Boyack, K. W. (2003) in *Annual Review of Information Science and Technology*, ed. Cronin, B. (Information Today, Inc., Medford, NJ), Vol. 37, pp. 179–255.

Visualization for constructing and sharing geo-scientific concepts

Alan M. MacEachren*, Mark Gahegan, and William Pike

GeoVISTA Center, Department of Geography, Pennsylvania State University, 302 Walker, University Park, PA 16802

Representations of scientific knowledge must reflect the dynamic nature of knowledge construction and the evolving networks of relations between scientific concepts. In this article, we describe initial work toward dynamic, visual methods and tools that support the construction, communication, revision, and application of scientific knowledge. Specifically, we focus on tools to capture and explore the concepts that underlie collaborative science activities, with examples drawn from the domain of human–environment interaction. These tools help individual researchers describe the process of knowledge construction while enabling teams of collaborators to synthesize common concepts. Our visualization approach links geographic visualization techniques with concept-mapping tools and allows the knowledge structures that result to be shared through a Web portal that helps scientists work collectively to advance their understanding. Our integration of geovisualization and knowledge representation methods emphasizes the process through which abstract concepts can be contextualized by the data, methods, people, and perspectives that produced them. This contextualization is a critical component of a knowledge structure, without which much of the meaning that guides the sharing of concepts is lost. By using the tools we describe here, human–environment scientists are given a visual means to build concepts from data (individually and collectively) and to connect these concepts to each other at appropriate levels of abstraction.

Scientific knowledge is dynamic. Its continuous evolution is marked by branches that diverge and converge and by conceptual frameworks that expand until they no longer support new insights, triggering dramatic reorganizations. In the earth sciences, perhaps the most poignant example is the theory of plate tectonics, originating in the early twentieth century with the work of Alfred Wegener, and eventually causing a massive reconceptualization of geological knowledge. Wilson (1) offers insight into this restructuring from a conceptual and philosophical perspective, and Giere (2) offers insight from a cognitive perspective. Although most changes in science are not as dramatic as those stimulated by the theory of plate tectonics, the concepts used by geologists, environmental scientists, and geographers to understand the Earth's complex systems and their interaction with human activities are nevertheless evolving as understanding evolves and as the needs of society change.

Information/geographic visualization can play a vital role in stimulating and communicating the evolution of conceptual structures. The case of plate tectonics provides a compelling example of the potential. In this case, visual representations influenced eventual acceptance of the theory (2). Specifically, the visual representations that provided evidence of tectonic activity interacted with geologists' different conceptualizations of the problem domain to produce both new concepts and new explanations for existing data (3).

Here, we focus on dynamic visual representations of conceptual frameworks that support (*i*) the process of knowledge construction and the application of that knowledge to scientific work and (*ii*) the connections between concepts in the mind and their instantiation in data. These visual representations can provide insight into the similarities and differences among scientific concepts held by a community of researchers. Moreover, visualization can serve as a vehicle through which groups of researchers share and refine concepts and even negotiate common conceptualizations. Our approach integrates geovisualization for data exploration and hypothesis generation, collaborative tools that facilitate structured discourse among researchers, and electronic notebooks that store records of individual and group investigation. By detecting and displaying similarity and structure in the data, methods, perspectives, and analysis procedures used by scientists, we are able to synthesize visual depictions of the core concepts involved in a domain at several levels of abstraction.

To contextualize our own work, and make the problem tractable, we focus on applicability of visual knowledge capture and representation methods for use in the science domain of human–environment interaction. Specifically, emphasis is on science work associated with the local human impacts of global environmental change. Knowledge about human–environment interaction is contextualized or "situated" by factors such as the places to which scientists direct their research, their aims, and their underlying theories. The structure of human–environment knowledge also depends on the choice of relevant datasets and methods and the scientist's experience applying them. Vulnerability to environmental changes, for example, is assessed (and possibly even conceptualized) quite differently in Massachusetts and Arizona. However, at certain levels of abstraction, some agreement among researchers in different locations about what constitutes vulnerability is essential for joint work. Geographic/information visualization can help collaborators construct and communicate knowledge structures that reflect the multidimensional connections among people, perspectives, data, and concepts at different conceptual scales.

The research we report is part of a science infrastructure project within which we are building a distributed collaboratory to support the work of four research sites that make up the Human Environment Regional Observatory (HERO) network. This developing collaboratory is an ideal "living laboratory" in which we can explore the construction of scientific knowledge and the role of visualization in enabling that construction. HERO collaborators are developing protocols to guide the collection of geospatial data for environmental monitoring and applying these protocols and data to problems such as assessing the vulnerability of local places to global environmental change

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

Abbreviation: HERO, Human Environment Regional Observatory.

*To whom correspondence should be addressed. E-mail: maceachren@psu.edu.

and the relationship between global environmental processes and local-scale land use/land cover change.

In the first section below we review previous research in both concept representation and the geographic/information visualization methods and tools on which our work builds. Then, we introduce the HERO problem context, review the concept of scientific collaboratories, and provide an overview of the HERO collaboratory. Next, we outline methods we have developed and implemented to support visually enabled concept building, sharing, and application. In the final section, we discuss future work.

Background

A critical component of science work involves developing, sharing, and comparing concepts and then applying those concepts to data to generate new knowledge. Within the HERO project, for example, a core concept is *vulnerability* of people and places to environmental change. This concept serves as a focal point that guides the process of posing research questions, collecting data, and producing communicable results. Vulnerability is also typical of many concepts in the social and environmental sciences; it is a complex, multifaceted, and context-dependent concept that has both everyday and scientific meaning. Despite the lack of apparent consistency in the meaning and application of this concept, the notion of vulnerability is embraced by many scientists in the environmental-change research community and is the subject of substantial research effort. Much of this research is geared toward developing measures of vulnerability that can be applied to specific places, but from these diverse local descriptions researchers attempt to synthesize the likely effects of global environmental change over large regions, even countries or continents. Moreover, the concept of vulnerability itself has broadened lately in response to recent world events; contemporary vulnerability measures attempt to account for risk from anthropogenic factors and from naturally occurring phenomena (4).

Our approach to understanding such concepts and facilitating their construction and use as a framework for science work draws on several research domains. Below, we briefly discuss two of these influences: research on concepts and their representation and research in information and geographic visualization relevant to visually enabled concept building and sharing.

Concepts and Concept Representation. The first driver for our research in developing tools to support knowledge construction is the rapidly evolving domain of concept representation. We define a "concept" as any abstract information resource that plays a role in scientific investigation. Scientific concepts do not necessarily include data or methods (although they may reflect and be constructed with data and methods). Rather, concepts may be categories, hypotheses, theories, or other constructions that help scientists organize their observations about the world. Thus, a person is not a concept, but the idea of a person is; the tangible form of an entity is given meaning in the world by the concepts we attribute to it. As we build tools to help scientists express and explore the concepts they use to describe the world, our task is to offer ways to structure and signify knowledge such that it can be communicated and reused most efficiently. The simplest form in which a concept might be signified could be a natural-language term, e.g., "plate tectonics," "grassland," or "water resources." Using the rules of language, scientists build more complex expressions of knowledge structures that link data, methods, theories, and other elements, which might be shared as journal articles. Our interest is not in such end products of scientific investigation, but in the process by which scientific knowledge evolves and in the development of tools to facilitate knowledge work in science.

The field of knowledge representation is concerned, in part, with computational aids to communication that reflect semantic relationships among concepts. Such representations can enable computational environments to visualize and reason with concepts by integrating knowledge structures within an individual's concept space and across multiple users and domains. Natural language is one medium for representing conceptual information, and in later sections we describe a visualization tool that helps reveal structure in natural discourse. However, it is (at present) difficult for systems that track the construction of concepts to reason with knowledge presented linguistically. As a result, our research relies on computational representations of knowledge and corresponding visual languages that allow inferences to be drawn more efficiently from complex bodies of scientific knowledge. These representations, as a complement to sharing knowledge through natural language, also support sharing and negotiation among scientists about the concepts that underpin joint work. We propose that the visually enabled concept representation and sharing methods we are developing will be particularly useful for asynchronous collaboration.

Typically, knowledge representation systems derive from first-order logic and its variants; popular examples include Prolog (5), Loom (6), and more recent developments such as frame logic (7). Despite (or perhaps because of) enforced decidability and consistency that can make knowledge representations effective for recording conceptual information computationally, most representational formats suffer from a syntax that is difficult for humans to create or parse. As a result, we favor notations such as diagrammatic reasoning tools (8) and conceptual graphs (9) that readily support visualization and both machine and human reasoning (it is possible to demonstrate equivalency between certain conceptual graph structures and predicate calculus or other logic). The concept visualization tools we describe later in this article couple a knowledge representation format based in description logic with these concept graphs; with these tools, researchers can diagram their thinking and have it stored as a set of description logic predicates that add to personal and communal knowledge bases.

Knowledge representation languages and the construction of ontologies that use them to describe features of the world have garnered substantial attention, not just in the computational sciences and artificial intelligence (e.g., ref. 10), but in the environmental and social sciences that are the focus of our present study (e.g., refs. 11–13). What is largely missing from this prior work, however, is consideration of how knowledge is generated, promulgated, revised, and retired. Knowledge representation implementations often focus on recording axioms about a domain without attempting to situate those axioms in the context of their creation or use. Environmental and social scientists grappling with the complexity of human–environment interaction are situated in a nexus of influences that includes their experience, their perspectives, and the places they study. These influences, rather than complicating the pursuit of objective truths, are fundamental to the nature of science work such that axiomatic knowledge cannot be cleanly separated from situated knowledge. Physicists see the world differently from geographers, not because there are different worlds to see, but because each community works within a historicity that gives concepts meaning in an evolving domain. This view of science is hermeneutic (developing out of ref. 14), embedding findings in a chain of interpretations, theories, models, methods, and measurements. If we wish to understand where ideas come from and where they go, we must incorporate references to situatedness in the representation and communication of scientific knowledge. Further, tools that support scientific knowledge representation must admit the situated-work practices of their potential users (15).

Knowledge representations, even those extended as we propose to include references to aspects of situation, do not by themselves achieve collaborative knowledge construction. Rather, knowledge representations must be embedded in tools

that help scientists communicate while preserving the context of their communication. To this end, many have described human–computer interaction as a conversation: with oneself, with one's collaborators, with one's descendents, with a machine (16, 17). We trade on this notion of a conversation as a means of helping researchers uncover the pedigree of shared ideas as they move from one scientist to another and from one time to another. This approach complements recent efforts in visualization of argumentation to support science work, discussed below (see ref. 18). Situated knowledge representations within collaborative software tools ground abstract ideas in a network of "conversations" across places, times, people, and perspectives. Occasionally, these conversations are explicit, and later we present results from visualizing Delphi method discussions as an example; often, however, they are not, and our work on electronic notebooks (see below) helps carry out implied conversations between researchers across place and time.

Visualizing (Geo)Concepts. The second driver of our research in developing tools to support knowledge construction is the combined domains of information/geographic visualization and diagrammatic reasoning. The information visualization community has developed a wide array of information exploration methods applicable to categorical data that can support interaction with scientific concepts. Many of these methods are designed to support hierarchical organization of information; examples include the cone tree (19), tree map (20), and the hyperbolic browser (21). Recent extensions include work by Robertson *et al.* (22) on the representation and exploration of multiple intersecting hierarchies and by Chen and Kuljis (23) and Fluit *et al.* (24) who focus explicitly on representation of knowledge domains. Other visual concept representation methods adopt a space-partitioning approach that assumes a single level; examples of these include extensions to Venn diagrams (25) and mosaic plots (26, 27). Still other methods focus on spatialization (28) of information, in general, text documents. Spatialization involves calculating the relationships among topics or concepts as distances in attribute space and "mapping" those relationships into a 2D or 3D space by using dimension reduction and cartographic representation methods (29, 30). Among the concept visualization techniques with a spatial metaphor, several have been developed and successfully used to construct or depict relationships among electronic resources [e.g., Fabrikant and Buttenfield (31), Havre *et al.* (32), and Miller *et al.* (33)].

When concepts involve geospatial components, as is common in human–environment interaction, developments in geovisualization, information visualization, and exploratory data analysis that support dynamically linked views, brushing, and focusing have considerable potential for adaptation to (geo)concept representation (34–37). One example is shown in Fig. 1. The view on the right is a standard choropleth map. It is dynamically linked to a graph browser (*Left*) that uses a minimum spanning tree approach to connecting places (counties, in this case) on the basis of their distance in multivariate attribute space (a spatialization of this attribute information). (This component is an extension of an open-source tool by Alex Shapiro called TOUCHGRAPH; see www.touchgraph.com.) In the example, the user clicked on Clearfield County to find other counties that are similar in attribute space.

The visualization methods noted above focus on helping users understand complex interrelationships within multivariate, often hierarchical, datasets. In general, they have not been directed to initial development (or acquisition) of concepts from individuals nor to sharing and comparing concepts among these individuals. However, research on diagrammatic reasoning environments has used visual techniques to facilitate development of knowledge by individuals and groups (8). That research has deep roots in domains such as legal argumentation, hypertext, and computer-

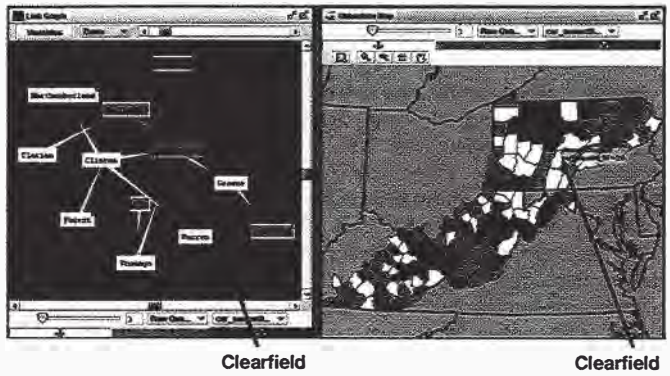


Fig. 1. Attribute space graph (*Left*) and linked map (*Right*). The attribute graph browser displays a combination of health (cancer mortality and success in diagnosis), demographic (census), and behavioral risk factors (smoking and obesity). Selection of Clearfield in the attribute space highlights counties in both attribute and geographic space that are similar to Clearfield in terms of all attributes. Most counties similar to Clearfield in this attribute space are nearby in geographic space.

mediated communication and has begun to produce robust tools and a rich body of research about how group thinking and negotiation can be enabled by visualization methods.

One example of diagrammatic reasoning tools with potential for application to scientific knowledge construction is provided by BELVEDERE, a software environment that supports the construction of diagrammatic representations of evidential relations (38). BELVEDERE enables remote collaboration and provides learners with shared workspaces for coordinating and recording their collaboration in scientific inquiry. It includes a visual representation language through which participants can build and share scientific arguments. Concepts that can be encoded include principle, theory, hypothesis, claim, and report; relationships include supports, explains, conflicts, justifies, and undercuts; and representations can be private, shared with all, or shared with a subgroup. Rinner (39) has conducted related work with place-based group knowledge building. His core idea involves implementing georeferenced annotation that is linked to a discussion forum focused on arriving at planning decisions. The resulting "argumap" is essentially a representation of the development of group knowledge and (perhaps) consensus. As outlined below, we are beginning to integrate a range of related visualization and visually enabled group work perspectives into tools for scientific knowledge work within the HERO project.

HERO Collaboratory

A core goal of the HERO project is to develop the technical and conceptual infrastructure to support long-term scientific research on local and regional human implications of global environmental change. A central part of our approach to achieving this goal is to develop a suite of methods and tools that facilitate synchronous and asynchronous joint work by small communities of scientists distributed around a network of sites across the United States. These methods and tools attempt to merge exploratory geovisualization tasks, during which concepts are constructed from data, with knowledge representation systems that capture the structure of relations between concepts, data, tools, and people.

HERO scientists are engaged in a variety of research programs, from developing protocols for data collection, through building theories and models to explicate multiscale processes of change, to developing policies to mitigate change. The mechanism used to make these methods and tools accessible to scientists and to enable joint knowledge construction in a

spatially and temporally distributed context is a scientific collaboratory (defined below).

Scientific Collaboratories: An Overview. The challenge of building national collaboratories was detailed in a 1993 National Research Council report (40). This report characterizes a collaboratory as a "center without walls, in which the nation's researchers can perform research without regard to geographical location—interacting with colleagues, accessing instrumentation, sharing data and computational resources, and accessing information from digital libraries." Considerable progress has been made toward the report goals (e.g., refs. 41–45). Emphasis thus far, however, has been on collaboratories that facilitate research in physical or medical sciences and on real-time data collection or control of experiments. Only limited progress has been made in application of the collaboratory concept to the study of human–environment interaction (46) or to fusing collaboratory concepts with work in collaborative geographic information systems (47) or collaborative geovisualization (48); see ref. 49 for more on map- and geographic information system-based collaboration. Also, little work has been done on application of knowledge representation methods, within collaboratories, to capture the semantic relationships between all the resources that a collaboratory may contain. Carroll *et al.* (50) and Chao *et al.* (51) describe the efforts of other science communities to use emerging knowledge management and portal technology to support knowledge construction in science.

The science establishment in the United States has recognized the need for what has been called "mega-collaboration" to address critical global problems (ref. 52; Zare was chair of the National Science Board at the time of this publication), and human vulnerability and responses to global environmental change is exactly the kind of problem where such megacollaboration is required. As noted by Finholt (44), barriers to interaction across distributed research sites will slow the construction and integration of the knowledge required to resolve challenging research questions. The goal of a distributed network, such as that being developed by HERO, is to bridge place and time by bringing researchers, the visual concept representation and sharing tools they use and the knowledge they build, to a single virtual environment.

Electronic Notebooks: A Vehicle for Acquiring, Constructing, and Sharing Knowledge. One component of the HERO collaboratory is a Web portal that integrates knowledge representation and information visualization tools in an electronic implementation of a traditional scientific notebook. Whereas paper notebooks were commonly used to record the development of an individual's ideas, our collaboratory notebooks are designed with the sharing and collective exploration of scientific information in mind. The notebook takes the form of an online workspace that gives investigators access not just to the digital data and tools they use (e.g., digital libraries, portals such as these are already becoming common) but to the abstract concepts constructed by using these data and methods. HERO workspaces provide a capacity to do more than just encode elements of scientific conversations that are easily "digitized." They also facilitate expressing and storing some of the reflection and reasoning that is usually tacit in the mind of the researcher. Rather than being stored in the form of a narrative, as might be common in a paper notebook, this reflection can be described visually through concept-graphing tools; the notebook system translates the resulting diagrams into a description logic-based knowledge representation language for storage and sharing.

Fig. 2 shows the home page of a user's workspace, providing access to the people he or she collaborates with, tasks that describe case studies or analysis procedures, concepts that define categories and ideas, data files used to create or reflect concepts, and online tools that can be used to visualize data and concepts. By using this portal system to describe elements of scientific investigations, researchers allow their electronic notebook to capture the evolution of their ideas and those of the communities of other users. Such a notebook allows common questions to be answered in new ways, and even some new questions to be asked, facilitating a dynamic process of concept and method development, extension, and application. For instance,

- Who first coined this concept?
- What data have been used to describe this concept?
- What alternative methods have been used to synthesize this concept?
- What concepts contain or are contained by this concept?
- Which individuals and groups have applied this concept?
- Do the reported aims of two individuals using the same concept agree?

Through the portal, HERO team members have access to personal workspaces that serve as a nexus for their own thinking and to group and community workspaces where common concepts can be synthesized from individual descriptions. A user can choose to make the contents of his or her workspace private or can make them available for crawlers to find in response to other users' queries. Group workspaces serve to collect points of agreement (or disagreement) between collaborating scientists (e.g., ref. 53), perhaps those working in a particular locality (such as a watershed) or on a specific problem. By contrast, community workspaces hold discipline-wide concepts that are broadly shared. In the context of human–environment research, a community notebook might define nationally or internationally agreed on concepts leading to shared protocols for vulnerability assessment or land use change analysis. Through time, concepts might migrate up and down such a hierarchy as they find or lose favor with their research community.

Visually Enabled Concept Building, Sharing, and Application

In this section, we present strategies for integration of exploratory geovisualization, information visualization, and diagrammatic reasoning methods and tools to support concept development, representation, and sharing. Specifically, we present some of our early steps toward achieving the separate aims of (*i*) building concepts from data and (*ii*) describing and communicating the relationships between concepts. We first discuss work focused on creating concepts to categorize natural features, deriving categories from data, and applying those categories to classification tasks. This work demonstrates the application of a range of integrated visual-computational methods to a subcomponent of the overall problem of concept building and sharing. We follow this with an outline of the initial set of methods and tools developed specifically to support concept building and sharing among HERO scientists. In the subsequent discussion section, we detail steps through which our portal approach will be extended to bridge the currently disconnected fields of visualization and knowledge sharing and apply them to work in a specific science domain.

From Concepts to Data and Back Again. Whereas it is often useful to impose *a priori* concepts on the analysis process, it is equally important in human–environment science to let concepts emerge by the combination of data, tools, and other situated aspects. Also, it is often necessary to mediate prior and emergent knowledge against each other (when theory and observation are not in accord). Indeed, it is at this interface that human–environment researchers regularly confront the dual problems of incomplete knowledge and incomplete data that characterize

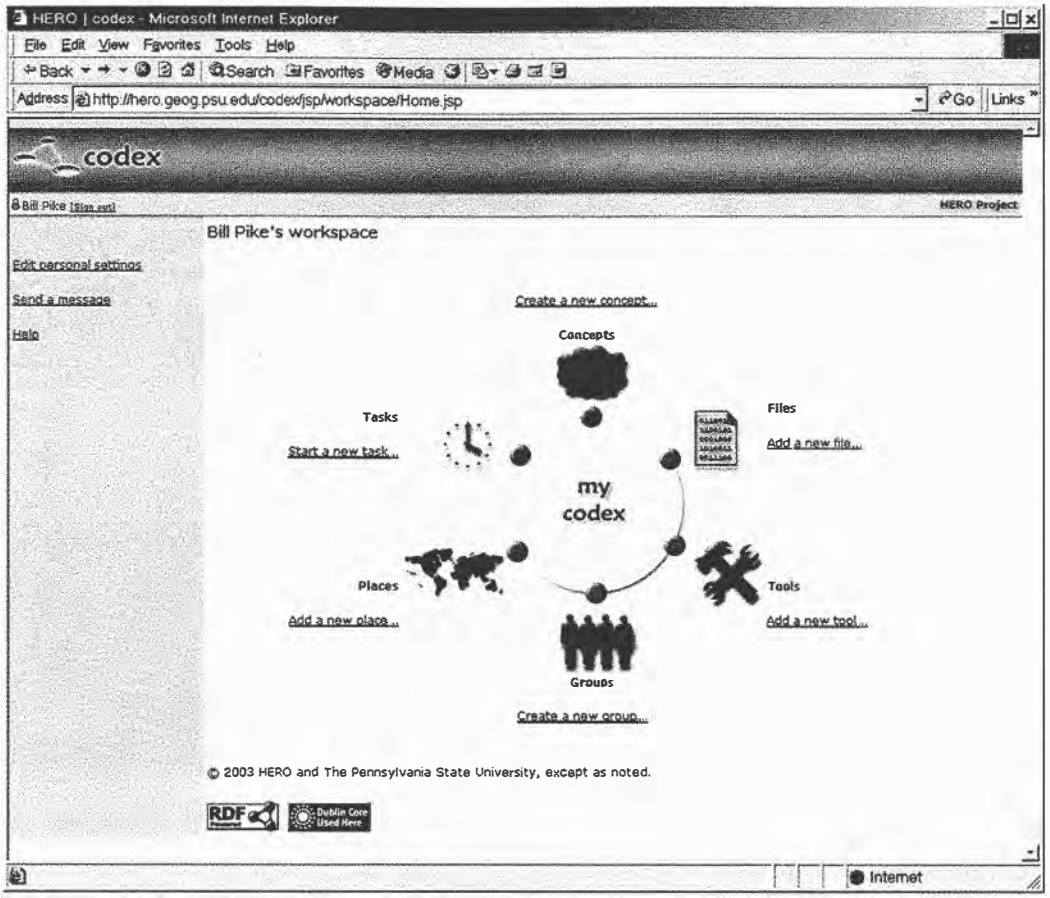


Fig. 2. Interface to knowledge representation and construction tools through a web portal.

their disciplines.[†] In this sense, human–environment science is both a descriptive and a discovery science; a person's understanding of concepts both helps to shape and is, in turn, shaped by interaction with data.

Many practitioners understand well that the creation of concepts is a compromise between their cognitive understanding of a problem and the emergent properties of the data. Therefore, concepts both impose structure on data and reveal the structure already present within the data (54). Our ultimate aim is to integrate these top-down and bottom-up approaches to knowledge application and knowledge construction. As noted above, this integration requires the fusion of two largely disparate research directions: the encoding and depiction of conceptual structures, such as situated, dynamic ontologies, representing what is known, and the support of data exploration and concept generation to test, refine, or derive conceptual structures, representing the discovery of new knowledge. At present, tools to support these activities are usually separated from each other with no means of interaction, but in practice activities at either end of this continuum are not isolated but intimately connected. As an example, consider the case of land cover and land use classification. Ontological tools that describe hierarchies of concepts (such as concepts associated with land use change that build on the Anderson land cover classification taxonomy) can offer sets of candidate categories from which a computational classifier might be trained or, conversely, exploring the clustering of sample points in attribute space might lead one to hypothesize suitable mental concepts to represent these points.

We have designed and are currently implementing and testing a suite of tools, developed in GeoVISTA *Studio* (www.geovistastudio.psu.edu; ref. 55) that connect the top-down processes of (*i*) defining and browsing concepts ontologically, (*ii*) selecting specific concepts to use in an analysis exercise, and (*iii*) operationalizing the concepts with classifiers with the bottom-up processes of (*iv*) data exploration to help formulate concepts from emergent structures in the data and (*v*) modification of the concepts, classifiers, or data used as a result of poor categories being produced from data (i.e., categories that do not align well with mental concepts or are not clearly differentiable in the data). Fig. 3 shows some of these tools, with arrows used to indicate some of their interactions within the process of geoscientific investigation (explained further in the legend to Fig. 3).

Ontological conceptualizations, as depicted at the upper left in Fig. 3, are created by individual or groups of researchers using a concept-graphing tool available through the HERO Web portal. This tool allows scientists to visually encode knowledge structures using conceptual graphing techniques. Users of this tool can produce diagrams to represent the relations between concepts or the process of an experiment or workflow. The example shown in Fig. 4 depicts one user's view of the concept of vulnerability to environmental change. Here, vulnerability is a product of three "subconcepts": exposure, sensitivity, and adaptation. Each of these concepts is in turn described by other concepts. All are linked together by using a set of relationships with defined semantics that allows the concept graph to be

[†]Although no shortage of available data exists, these data do not completely describe one's objects of study. Just as concepts merely refer to more abstract representations in the mind, data are a proxy for the phenomena they are intended to measure. For instance, there is no objective measurement for the concept of "vulnerability"; there are only other phenomena, such as flood frequency or demographics, that may be measured (and even these, incompletely).

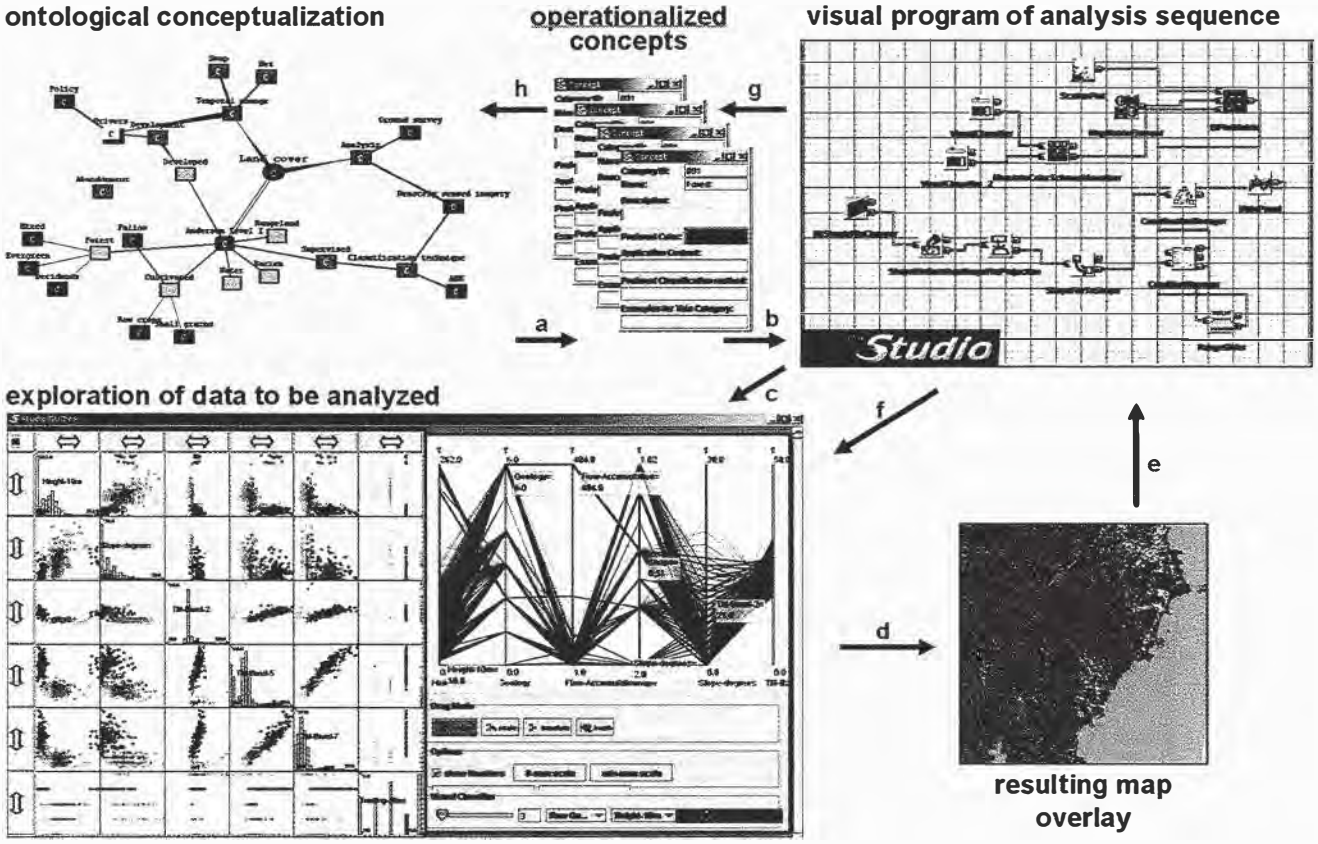


Fig. 3. Overview of coordinating bottom-up and top-down approaches to analysis. (a) Concepts to be used in an analysis are extracted from the ontology and held in an experimental notepad. (*b*) Design for the experiment is constructed by using the *Studio* visual programming utility. (*c*) The data are analyzed for emergent structures and relationships that can be utilized and for errors and unhelpful attributes that possibly should be removed. (*d*) The experiment produces a result set of categories, held intentionally as pieces of a classifier model and extensionally as a map or dataset. (e) Problems with the result can cause the experimental design to be changed. (*f*) Problems with the result might lead to a reexploration of the data. (*g*) Problems with the result might cause the user to modify the concepts being utilized. (*h*) Modified concepts can be inserted into the ontology, leading to a modified ontology.

decomposed into a set of concept definitions stored in description logic. Some nodes on this graph represent data files, and the links between these nodes and other concepts suggest what observations might be used to describe abstract concepts.

Visualizing Structure in Scientific Discourse. In this subsection we return to the idea, introduced above, of natural language as a knowledge representation. One component of the HERO collaboratory available through the Web portal is a tool for conducting online Delphi method activities. The Delphi method (56) is a multiparticipant technique for eliciting and refining expert belief and is used by HERO researchers to synthesize core concepts involved in phenomena such as vulnerability to environmental change. Through Delphi exercises, we enable another type of scientific "conversation" to be performed. By using natural language processing techniques, the key themes in Delphi discussions can be extracted from the content of participants' postings; these themes are then compared against a lexical database that helps organize them into conceptual hierarchies, which participants can browse to navigate a discussion or to summarize the important ideas in the science domain under discussion. Fig. 5 shows a graph browser displaying concept relationships from a Delphi discussion on vulnerability (the browser uses the same underlying technology as that in Fig. 1).

The key themes in the discussion have been automatically aggregated to higher levels of abstraction. In this case, the concepts emerged from the text through bottom-up processing, but are being viewed by this user in a top-down fashion. At the center of the graph are the most general representations of the concepts associated with vulnerability, as expressed in the discourse. Some nodes have been expanded to show increasingly specific representations of those concepts and can continue being expanded until the actual terms used in the discussion appear. This conceptual graph is populated exclusively with "kind-of" relationships between concepts, yet it demonstrates a technique useful for extracting concepts from text data. Ultimately, a Delphi concept map could be produced to show a set of domain concepts extracted from text discussions or journal articles; these concepts can then be linked to data and geovisualization tools that would help describe them further.

Discussion (Future Work): Visually Enabled Knowledge Work

As noted above, this article provides just a sketch of a comprehensive conceptual approach we are developing for enabling and understanding the process of concept construction in human–environment science. Further work needs to be directed toward formalizing this approach, extending methods for concept visualization, integrating visualization with groupware to enable visual support for group thinking, and applying the results in the living laboratory of the HERO project. Specific next steps are detailed below.

Formalizing the Approach. The framework for formalizing our approach to concept representation is based on extensions to the DARPA Agent Markup Language + Ontology Inference Layer

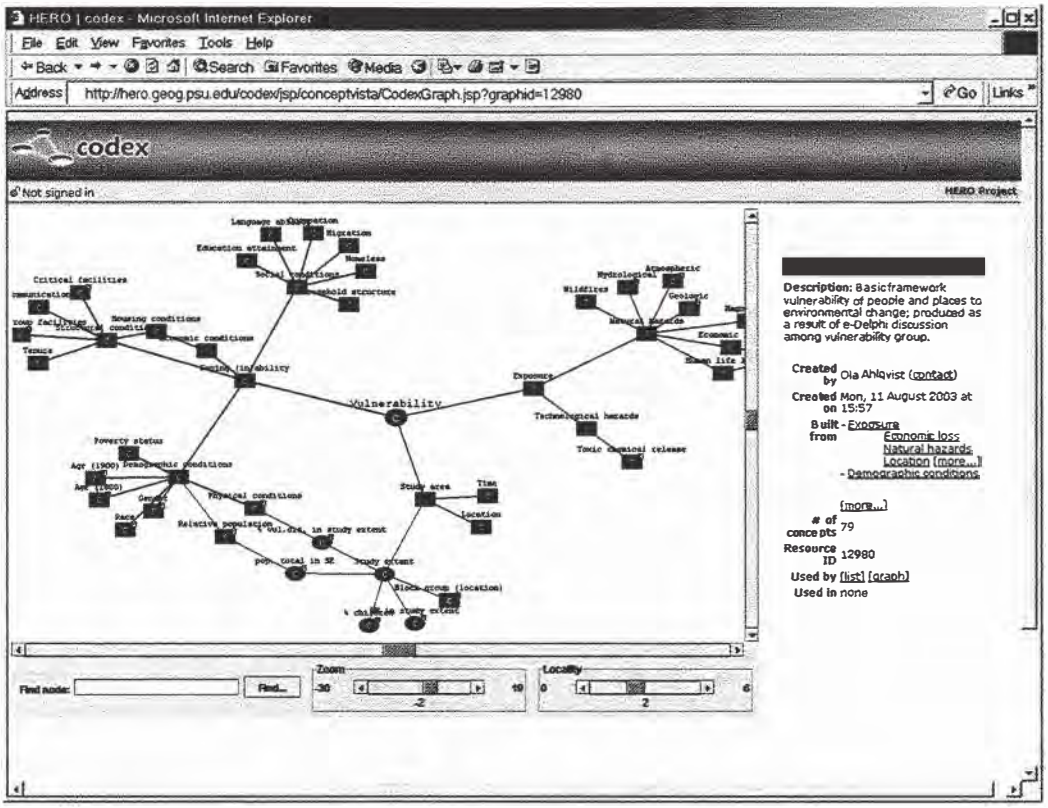


Fig. 4. A concept graph that depicts a HERO researcher's conceptualization of vulnerability. The graph allows concepts, data, and tools to be linked in visual descriptions of the research process.

(DAML+OIL) markup language, which can be expressed in XML. These extensions combine a frame-based syntax with description logic inference rules. Semantic information is stored in a discrete and portable format that enables collaborators to share concepts easily through the HERO knowledge portal. The framework we have implemented thus far supports construction of concepts that refer to one or more ontologies, linked concept networks in which any concept can become an attribute of another concept and ontologies that can be individual or shared. Ongoing work involves coupling the formalization of concepts and concept structures to visual tools that support both direct construction of individual concepts (by individuals) and collaboration among individuals to develop shared concepts and ontologies of which they are a part.

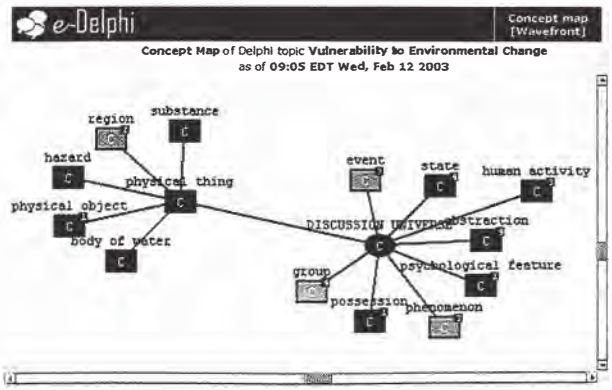


Fig. 5. Conceptual graph showing top-down view of concepts extracted from online discussion on vulnerability.

Extending Visualization Methods to Better Support Concept Building, Representation, and Comparison. One goal here is to develop visual-computational methods that support comparison of concept maps. These methods will allow scientists to compare their own representation of a concept with that of other individual scientists or with the group view(s) derived from Delphi discussions. Such comparisons can reveal points of tension within a community's view of a domain and help to clarify distinctions between a novel extension to a concept and the accepted (group) view. Computational comparison will include graph-similarity measures (e.g., maximum common subgraph) for evaluating overlap between multiple ontologies in DAML.

Integrating Visualization with Groupware. A related goal is to draw on the range of recent developments in methods for visually enabled group work, diagrammatic reasoning, and argument visualization and fuse them by exploratory visualization methods to provide a flexible environment to support group knowledge building.

Tailor the Methods and Tools for Specific HERO Activities. As a proof-of-concept test for methods and tools, we will adapt them for use by HERO team members in individual and community concept development focused on the concepts of vulnerability, water resource management, and land use change and the related concepts from which each is composed.

Each of the objectives above is being advanced through the work of 12 student researchers who are part of the HERO Research Experience for Undergraduates Site. These

students will apply the initial tools to the problem of understanding "sensitivity" of local water resources to environmental change.

This work was supported by National Science Foundation Grants BCS-9978052, BCS-0113030, and BCS-0219025 and by the U.S. Geological Survey.

1. Wilson, J. T. (1968) *Proc. Am. Philos. Soc.* **122,** 309–320.
2. Giere, R. N. (1988) *Explaining Science: A Cognitive Approach* (Univ. of Chicago Press, Chicago).
3. MacEachren, A. M. (1995) *How Maps Work: Representation, Visualization and Design* (Guilford Press, New York).
4. Muntz, R. R., Barclay, T., Dozier, J., Faloutsos, C., MacEachren, A. M., Martin, J. L., Pancake, C. M. & Satyanarayanan, M. (2003) *IT Roadmap to a Geospatial Future: Report of the Committee on Intersections Between Geospatial Information and Information Technology* (Natl. Acad. Press, Washington, DC).
5. Colmerauer, A. & Roussel, P. (1993) *SIGPLAN Notices* **28,** 37–52.
6. MacGregor, R. (1994) in *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, eds. Hayes-Roth, B. & Korf, R. E. (Am. Assoc. for Artificial Intelligence, Seattle), pp. 213–220.
7. Kifer, M., Lausen, G. & Wu, J. (1995) *JACM* **42,** 741–843.
8. Kirschner, P. A., Shum, S. J. B. & Carr, C. S. (2003) *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-making* (Springer, London).
9. Sowa, J. (2000) *Knowledge Representation: Logical, Philosophical, and Computational Foundations* (Brooks/Cole, Pacific Grove, CA).
10. Guarino, N. (1997) *Int. J. Hum. Comput. Stud.* **46,** 293–310.
11. Doel, M. A. (2001) *Environ. Plann. D Soc. Space* **19,** 555–572.
12. Fonseca, F. T., Egenhofer, M. J. & Agouris, P. (2002) *Trans. in GIS* **6,** 231–257.
13. Frank, A. U. (2001) *Int. J. Geogr. Inf. Sci.* **15,** 667–678.
14. Gadamer, H.-G. (2003) *Truth and Method* (Continuum, New York).
15. Schultze, U. & Boland, R. J. (2000) *J. Strategic Inf. Syst.* **9,** 193–212.
16. Nake, F. & Grabowski, S. (2001) *Knowledge-Based Syst.* **14,** 441–447.
17. Winograd, T. & Flores, F. (1986) *Understanding Computers and Cognition* (Ablex, Norwood, NJ).
18. Shum, S. B., Uren, V., Li, G., Domingue, J. & Motta, E. (2003) in *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, eds. Kirschner, P. A., Shum, S. J. B. & Carr, C. S. (Springer, London), pp. 185–204.
19. Robertson, G. G. (1991) in *Proceedings of CHI 91: Conference on Human Factors in Computing Systems*, eds. Robertson, S. P., Olson, G. M. & Olson, J. S. (Association of Computing Machinery, New York), pp. 189–194.
20. Johnson, B. & Shneiderman, B. (1991) in *Proceedings of IEEE Visualization'91*, eds. Nielson, G. M. & Rosenblum, L. J. (IEEE, Piscataway, NJ), pp. 284–291.
21. Lamping, J., Rao, R. & Pirolli, P. (1995) in *Proceedings of CHI 95: Human Factors in Computing Systems*, eds. Katz, I. R., Mack, R., Marks, L., Rosson, M. B. & Nielson, J. (ACM Press, New York), pp. 401–408.
22. Robertson, G., Cameron, K., Czerwinski, M. & Robbins, D. (2002) *J. Inf. Visualization* **1,** 50–65.
23. Chen, C. & Kuljis, J. (2003) *J. Am. Soc. Inf. Sci. Technol.* **54,** 435–446.
24. Fluit, C., Horst, H. t. & van der Meer, J. (2002) in *On-To-Knowledge: Content-Driven Knowledge Management Tools Through Evolving Ontologies,* EU-IST Project IST-1999-10132, Report, Dec. 9, 2002, (Commission of the European Communities, Amsterdam, The Netherlands).
25. Marshall, R. J. (2001) *Stat. Med.* **20,** 1077–1088.
26. Hartigan, J. A. & Kleiner, B. (1984) *Am. Statistician* **38,** 32–35.
27. Friendly, M. (1994) *J. Am. Stat. Assoc.* **89,** 190–200.
28. Kuhn, W. & Blumenthal, B. (1996) *Spatialization: Spatial Metaphors for User Interfaces* (ACM Press, New York).
29. Wise, J., Thomas, J., Pennock, K., Lantrip, D., Pottier, M., Schur, A. & Crow, V. (1995) in *Proceedings of IEEE Symposium on Information Visualization 1995*, eds. Gershon, N. & Eick, S. (IEEE, Piscataway, NJ), pp. 51–58.
30. Skupin, A. (2000) in *Proceedings of IEEE Symposium on Information Visualization 2000 (InfoVis 2000)*, eds. Roth, S. & Keim, D. (IEEE, Piscataway, NJ), pp. 91–98.
31. Fabrikant, S. I. & Buttenfield, B. P. (2001) *Ann. Assoc. Am. Geogr.* **91,** 263–280.
32. Havre, S., Hetzler, B. & Nowell, L. (2000) in *Proceedings of IEEE Symposium on Information Visualization 2000 (InfoVis 2000)*, eds. Roth, S. & Keim, D. (IEEE, Piscataway, NJ), pp. 115–123.
33. Miller, N., Wong, P., Brewster, M. & H., F. (1998) in *Proceedings of IEEE Symposium on Information Visualization 1998*, eds. Wills, G. & Dill, J. (IEEE, Piscataway, NJ), pp. 189–196, 532.
34. Fredrikson, A., North, C., Plaisant, C. & Shneiderman, B. (1999) in *Proceedings of the Workshop on New Paradigms in Information Visualization and Manipulation (NPIVM '99)* (ACM Press, New York), pp. 26–34.
35. Andrienko, G. L. & Andrienko, N. V. (1999) *Int. J. Geogr. Inf. Sci.* **13,** 355–374.
36. Gahegan, M., Harrower, M., Rhyne, T.-M. & Wachowicz, M. (2001) *Cartogr. Geogr. Inf. Sci.* **28,** 29–44.
37. MacEachren, A. M., Hardisty, F., Dai, X. P. & Pickle, L. (2003) *Commun. ACM* **46,** 59–60.
38. Suthers, D. (1999) in *Proceedings of the 32nd Hawaii International Conference on System Sciences 1999*, eds. El-Renwini, H. & Helal, S. (IEEE Computer Society Press, Los Alamitos, CA).
39. Rinner, C. (2001) *Environ. Plann. B Plann. Design* **28,** 847–863.
40. Cerf, V. G., Cameron, A. G. W., Lederberg, J., Russell, C. T., Schatz, B. R., Shames, P. M. B., Sproull, L. S., Weller, R. A. & Wulf, W. A. (1993) *National Collaboratories: Applying Information Technology for Scientific Research* (Natl. Acad. Press, Washington, DC).
41. Kouzes, R. T., Myers, J. D. & Wulf, W. A. (1996) *Computer* **29,** 40–46.
42. Henline, P. (1998) *Interactions* **May/June,** 66–72.
43. Olson, G. M., Atkins, D. E., Clauer, R., Finholt, T. A., Jahanian, F., Killeen, T. L., Prakash, A. & Weymouth, T. (1998) *Interactions* **May/June,** 48–55.
44. Finholt, T. A. (2001) *Annu. Rev. Inf. Sci. Technol.* **36,** 73–108.
45. Olson, G. M., Malone, T. W. & Smith, J. B. (2001) *Coordination Theory and Collaboration Technology* (Lawrence Erlbaum Associates, Mahwah, NJ).
46. Kuhlman, K. M., Soffer, A. & Foresman, T. W. (1997) in *Second IEEE Metadata Conference* (IEEE, Piscataway, NJ).
47. Churcher, C. & Churcher, N. (1999) *Trans. Geogr. Inf. Syst.* **3,** 23–30.
48. Rhyne, T. M. & Fowler, T. (1998) in *ACM SIGGRAPH 98 Course Notes*, no. 35 (ACM Press, New York).
49. MacEachren, A. M. (2001) *Prog. Hum. Geogr.* **25,** 431–444.
50. Carroll, J., Rosson, M.-B., Dunlap, D. & Isenhour, P. (2003) in *Proceedings of the 36th Hawaii International Conference on System Sciences 2003*, ed. Sprague, R., Jr. (IEEE, Piscataway, NJ), pp. 120–129.
51. Chau, M., Chen, H., Qin, J., Zhou, Y., Sung, W. K., Chen, Y., Qin, Y., McDonald, D., Lally, A. & Landon, M. (2002) in *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'02)* (ACM Press, New York), p. 373.
52. Zare, R. N. (1997) *Science* **275,** 1047.
53. Harvey, F. & Chrisman, N. (1998) *Environ. Plann. A* **30,** 1683–1694.
54. Anderberg, M. R. (1973) *Cluster Analysis for Applications* (Academic, New York).
55. Gahegan, M., Takatsuka, M., Wheeler, M. & Hardisty, F. (2002) *Comput. Environ. Urban Syst.* **26,** 267–292.
56. Turoff, M. & Hiltz, S. (1995) in *Gazing into the Oracle: The Delphi Method and its Application to Social Policy and Public Health*, ed. Ziglio, E. (Kingsley, London).

Mapping topics and topic bursts in PNAS

Ketan K. Mane and Katy Börner*

School of Library and Information Science, Indiana University, 10th Street and Jordan Avenue, Bloomington, IN 47405

Scientific research is highly dynamic. New areas of science continually evolve; others gain or lose importance, merge, or split. Due to the steady increase in the number of scientific publications, it is hard to keep an overview of the structure and dynamic development of one's own field of science, much less all scientific domains. However, knowledge of "hot" topics, emergent research frontiers, or change of focus in certain areas is a critical component of resource allocation decisions in research laboratories, governmental institutions, and corporations. This paper demonstrates the utilization of Kleinberg's burst detection algorithm, co-word occurrence analysis, and graph layout techniques to generate maps that support the identification of major research topics and trends. The approach was applied to analyze and map the complete set of papers published in PNAS in the years 1982–2001. Six domain experts examined and commented on the resulting maps in an attempt to reconstruct the evolution of major research areas covered by PNAS.

Maps depicting the structure and evolution of scientific fields are also called knowledge domain visualizations (KDVs) (1). They are a special kind of information visualization (2) that exploits powerful human vision and spatial cognition to help humans mentally organize and electronically access and manage large complex information spaces. Unlike scientific visualizations, KDVs are created from data that have no spatial reference, such as sets of publications, patents, or grants.

KDVs use sophisticated data analysis and visualization techniques to objectively identify major research areas, experts, institutions, grants, publications, journals, etc., in a domain of interest. They can be used to gain an overview of a knowledge domain; its homogeneity, import–export factors, and relative speed; to track the emergence and evolution of topics; or to help identify the most productive as well as new research areas. Benefits of KDVs include reducing visual search time, revealing hidden relations, displaying data sets from several perspectives simultaneously, facilitating hypothesis formulation, serving as effective means of communication, and prompting users to think in new ways about document data. Today, KDVs are typically generated semiautomatically from rather small static data sets and for a specific knowledge domain and information need.

This paper presents a way to generate co-word association maps of major topics based on highly frequent words and words with a sudden increase in usage, a phenomenon called "burst" (3). A large-scale data set comprising the complete set of 47,073 papers published in the PNAS in the years 1982–2001 is used for demonstration. We describe the identification of highly frequent and bursty words, the analysis of the most important correlations among those words, and the generation and interpretation of a 2D layout showing major research topics and their dynamics.

Tracking the Evolution of Major Topics

To identify major words or topics covered in PNAS, we first selected the top 10% of the most highly cited documents for each of the 20 years. This is common practice, because papers with few citations are assumed to have less impact, and most algorithms simply cannot handle very large amounts of data. The least-cited paper in this set received 14 citations.

The next step is the identification of major sources for potential topic words. Biologists are well aware that titles and key words are used for indexing. Hence, they tend not to use words that occur in the title as key words and vice versa. Therefore, paper titles and key words were selected for the subsequent analysis. Two types of key words exist: Institute for Scientific Information (ISI) key words,[†] which come from the author or publisher, or titles of cited papers and MEDLINE's controlled vocabulary, also called MeSH terms. No ISI key words are available for papers published before 1991. MeSH terms have been joined to ISI records by using the procedure in Boyack (4). Papers without titles were excluded from the analysis, resulting in 4,699 papers. From those papers, a total of 34,299 unique potential topic words were extracted.

To determine the trends of word usage over time, the top 10 most frequent and most meaningful words were then selected in collaboration with domain experts. Those words, in order of decreasing frequency, are human, animal, mice, molecular sequence data, genes, expression, RNA, DNA, cell line, and cloning. Excluded from the top frequency list were support, U.S. Gov't, non-U.S. Gov't, P.H.S., receptors, cells, rats, amino acid sequence, base sequence, and cultured.

Fig. 1 shows the frequency count for all 10 words for the 20-year time period. Clearly visible is the introduction of new terms as well as the increase or decrease in the usage of certain words and the influence of biotechnology events.[‡] For example, human studies are steadily increasing and are bursting at the start of the Human Genome Project in 1988. Research on animal and mice shows a similar trend. Research on genes, DNA, and RNA is strongly coupled and shows similar upward trends.

As exemplified in Fig. 1, several research trends have occurred in parallel with word frequencies. For example, PCR technology, conceived in 1983 and combined with reverse transcription, was fully developed by 1986 and commonly used for gene expression experiments. Cell-line research was carried out during the 1982–1991 time period. It declined in later years as focus shifted from research to application development. Molecular sequence data research received a boost with the sanction of funds for the Human Genome Project. Its importance decreased after 1994 as it became routine in genetic studies. In 1991, work on the expression rate of DNA to protein conversion (rather than just the transcription process) ignited. As more data on molecular

---

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

Abbreviation: KDV, knowledge domain visualization.

*To whom correspondence should be addressed. E-mail: katy@indiana.edu.

[†]These data are extracted from *Science Citation Index Expanded* [Institute for Scientific Information, Inc. (ISI), Philadelphia, PA; Copyright ISI]. All rights reserved. No portion of these data may be reproduced or transmitted in any form or by any means without the prior written permission of ISI.

[‡]Biotechnology timeline, Biotechnology Institute (www.biotechinstitute.org/pdf/bio_timeline.pdf).

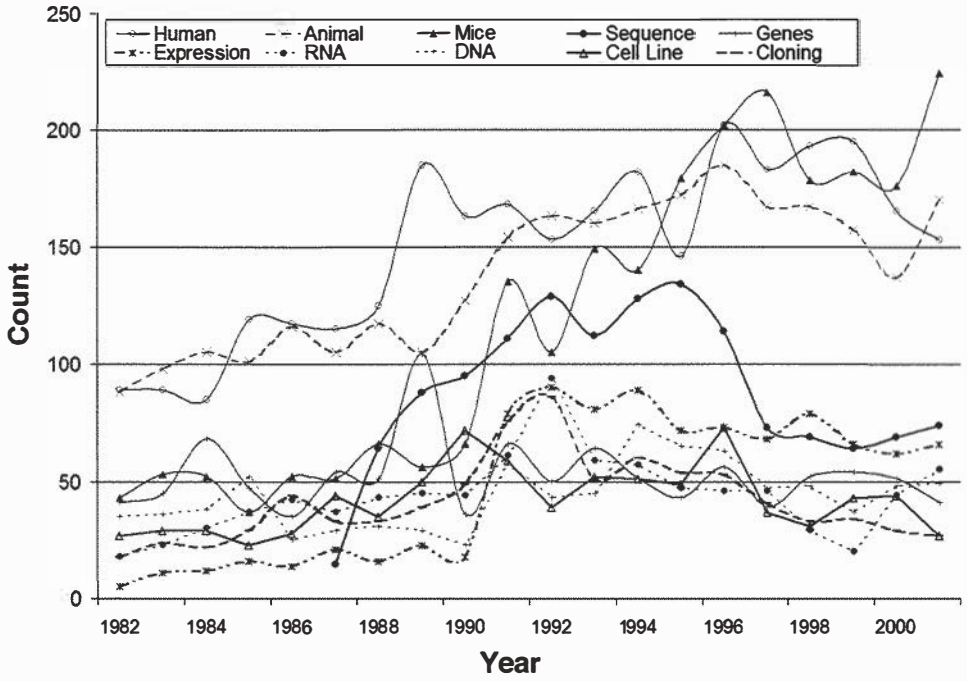


Fig. 1. Frequency count for the most frequently used words in the top 10% of most highly cited PNAS publications from 1982 to 2001.

sequence, expression rate, etc., became available, research toward developing molecular models and protein modeling became relevant. Cloning research had its first breakthrough result with the successful cloning of a sheep in 1997. Molecular cloning approaches were quickly adopted in subsequent years.

In addition to word frequency, the dynamics of topic usage were also examined. Kleinberg's burst detection algorithm (3) was applied to identify topics that experience a sudden increase in usage, also called burst. The burst detection algorithm provides a formal model for the robust and efficient identification of word bursts. Using a finite-state automaton, bursts in streams of words correspond to state transitions. The algorithm outputs the start and end time of a burst as well as its strengths for each word. Some words in the PNAS data set experience multiple bursts. In the top 10% of the most highly cited PNAS publications, there were 1,027 unique words, of which 991 had at least one burst, and 34 of which had two bursts. Exactly two words, "comparative study" and "dna primers," bursted on three occasions but are not among the highly frequent words.

Mapping the Co-Occurrence Space of Topics and Topics Bursts

After analyzing the frequency occurrence of all 34,299 unique words for each of the 20 years as well as their burstyness, we next chose to identify and map the relationships among major topics.

Co-word occurrence analysis is a content analysis technique that can be used to identify the strength of associations between words based on their co-occurrence in the same document (5). Words that appear together often will have a strength closer to 1, and words that never appear together, a strength of 0. Although co-word spaces are typically generated based only on highly frequent words, the work presented here is unique, because it also accounts for word burstyness.

To begin, we computed the intersection of the highest frequency and most highly bursting word sets and selected the first 50 for further analysis. Interestingly, there was a rather low correlation among the frequency of words and their burstyness. In the particular example discussed here, it took 742 most frequent words and 874 most bursty words to get an intersection of 50 words.

The co-word analysis was conducted for those 50 words and the set of 4,699 documents. The resulting co-occurrence frequency matrix was normalized by using Salton's cosine coefficient (6), where each word pair co-occurrence is defined as the ratio of their co-occurrences and the product of the square root of the respective word occurrences within the document set. Interestingly, the original nonnormalized co-occurrence matrix resulted in more meaningful maps, as judged by domain experts, and will be used in the subsequent discussion.

The nonnormalized co-occurrence matrix has 1,082 nonzero entries characterizing the complex co-occurrence relationships among the 50 words. To reduce this number to the most meaningful relationships, the pathfinder network scaling algorithm (PFNet) (7) was applied. The PFNet algorithm relies on triangle inequality to eliminate unnecessary links. Given two paths (sequence of links) in a network that connects two nodes, the path that has a greater weight as defined via the Minkowski metric is preserved. It is assumed that a path with a greater weight better captures the interrelationship between two nodes, and that alternative paths with less weight are redundant or even counterintuitive and should be pruned from the network. Two parameters, r and q, influence the topology of a pathfinder network. The r parameter specifies the weight of a path based on the Minkowski metric. The q parameter defines the number of links in alternative paths (i.e., the length of a path) up to which the triangle inequality must be maintained. A network of n nodes can have a maximum path length of $q = n-1$. With $q = n-1$, the triangle inequality is maintained throughout the entire network. A detailed explanation of PFNet as well as another application of it is reported in Chen (8).

Running PFNet with $q = n-1 = 49$ and infinite r results in 62 nonzero entries and hence a very sparsely connected network of 50 topic nodes and 62 edges. The many lattice-like subgraphs in this network reveal only part of the complex relationships among the major topics. Based on expert feedback, we selected the network with 80 edges that was generated by using the parameter values $r = 6$ and $q = 49$.

Subsequently, the topic word co-occurrence network was laid out in two dimensions for visual examination. The layout de-

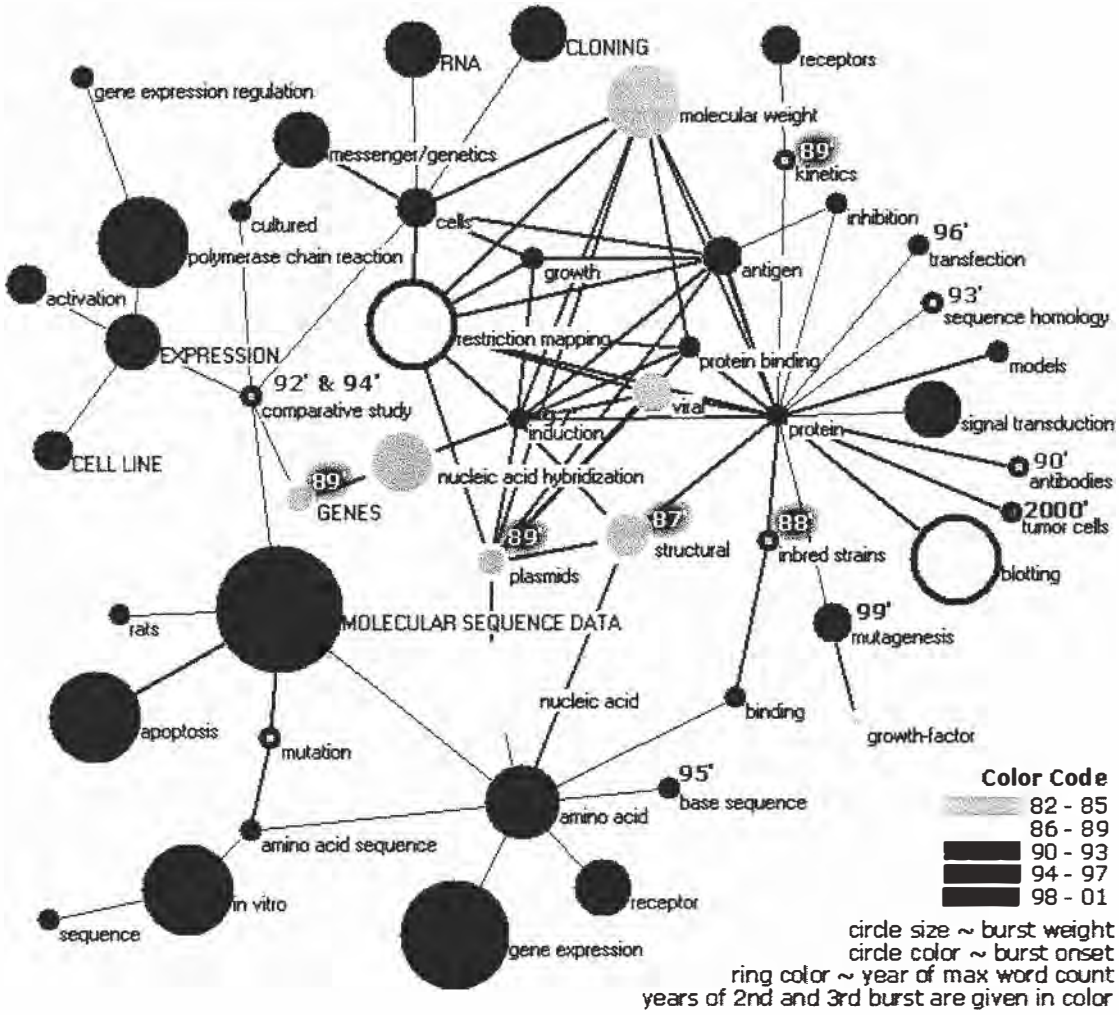


Fig. 2. Co-word space of the top 50 highly frequent and bursty words used in the top 10% most highly cited PNAS publications in 1982–2001.

picted in Fig. 2 was generated by using the Fruchterman–Reingold 2D graph layout algorithms (9), a more efficient version of the original spring embedding algorithm developed by Eades (10). Each node in the network represents one of the 50 highly frequent and bursting words. Note that the words human, animal, mice, and DNA discussed in Fig. 1 do not burst and hence are not included in Fig. 2. The size of the node circle corresponds to the maximum burst level this word achieved. Color coding is used to denote the years in which the word was used most often as well as the year of the maximum burst. Five time durations and respective colors were used: 1982–1985, green; 1986–1989, yellow; 1990–1993, orange; 1994–1997, red; and 1998–2001, black. The year of the maximum frequency and the starting year of the first burst of this word were decoded by circle border colors and inner circle area colors, respectively. For example, the word molecular sequence data, represented by a circle of rather large size with an orange inner area and a red ring, showed the highest, rather large burst between 1990 and 1993 and had a high frequency of usage in the later years 1994–1997.

Edge thickness is proportional to the number of word co-occurrences. Protein and models or cells and growth are co-occurring frequently in the selected publication data set.

The evaluation and interpretation of the resulting map are rather difficult, because there are very few people who are familiar with the diverse research results reported in the original data set (4,699 highly cited PNAS papers published over a 20-year time span). However, the graph visualization shown in Fig. 2 was examined by six biologists, and their interpretation of the map is summarized below.

Over the last 20 years, biological research has experienced enormous growth and also diversification. The graph shown in Fig. 2 semantically interrelates and chronologically links diverse fields of biological research. Four major areas can be identified that are interlinked with the middle-oldest area of research. The top left subnetwork is related to expression profiling and genomics research, top right topics deal with protein research, right bottom topics are linked to cancer research, and the ones in the left bottom relate to molecular sequence studies.

In the early 1980s, the primary research foci were structural properties of biological entities such as cells, genes, etc. This was followed by a phase of research in kinetics and the study of the mutation behavior of genes. Toward the early 1990s, in conjunction with the start of the Human Genome Project, the research paradigm shifted toward sequence data studies. During this time period, molecular sequence data, amino acid sequences associated with the genome project, rose to prominence. Major funding via the Human Genome Project also brought together several interconnected research areas primarily dealing with cloning, PCR, and gene expression depicting cloning studies. These experiments are an extension of prior studies on plasmids, genes, and nucleic acid hybridization. In later years, research

concentrated on apoptosis, signal transduction, activation, and cells linked by cell signaling pathways for programmed cell death, a key area of cancer research. The increase in computing power facilitated extensive research in modeling research, leading in turn to an increased understanding about the folding patterns of proteins. In 2000, the human genome sequence was completed, and investigations now concentrate on protein research.

Summary

Here we have demonstrated an objective computational approach to analyzing the structure and evolution of a research domain. To our knowledge, this is the first attempt to map the co-word space of highly frequent and bursty words. The resulting visualization depicts 50 major topics and topic bursts in PNAS and their evolution over a 20-year time frame.

Problems of dimensionality reduction for generating plots of high-dimensional data sets were tackled by using threshold values to select a representative document and unique word set as well as the application of pathfinder network scaling to capture major associations among words.

The resulting visualizations were examined and interpreted by a number of domain experts, demonstrating their readability and practical value for the identification of topics, major trends, and research frontiers as well as hinting at their value as a knowledge management tool for researchers, companies, funding agencies, and society.

We thank Anne Prieto, Don G. Gilbert, Sun Kim, Keith G. Ngolley, Kranthi Varala, and Claire Nisonger for insightful comments on the interpretation of the generated maps and Margaret Swan for proofreading this paper. The KNOT tools for Pathfinder Network Analysis (http://interlinkinc.net) and PAJEK (11) were used in the presented analysis. This work is supported by National Science Foundation CAREER Grant IIS-0238261 and National Science Foundation Grant DUE-0333623.

1. Börner, K., Chen, C. & Boyack, K. (2003) in *Annual Review of Information Science and Technology*, ed. Cronin, B. (Information Today/American Society for Information Science and Technology, Medford, NJ), Vol. 37, pp. 179–255.
2. Card, S., Mackinlay, J. & Shneiderman, B. (1999) *Readings in Information Visualization: Using Vision to Think* (Morgan Kaufmann, San Francisco).
3. Kleinberg, J. M. (2002) in *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York), pp. 91–101.
4. Boyack, K. W. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5192–5199.
5. Callon, M., Law, J., & Rip, A. (1986) *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World* (Macmillan, London).
6. Salton, G. (1989) *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer* (Addison–Wesley, New York).
7. Schvaneveldt, R. W. (1990) *Pathfinder Associative Networks: Studies in Knowledge Organization.* (Ablex, Norwood, NJ).
8. Chen, C. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5303–5310.
9. Fruchterman, T. M. J. & Reingold, E. M. (1991) *Software Pract. Exp.* **21,** 1129–1164.
10. Eades, P. (1984) *Cong. Numer.* **42,** 149–160.
11. Batagelj, V. & Mrvar, A. (1998) *PAJEK–A Program for Large Network Analysis Connections* **21,** 47–57.

Crossmaps: Visualization of overlapping relationships in collections of journal papers

Steven A. Morris* and Gary G. Yen

Electrical and Computer Engineering, Oklahoma State University, 202 Engineering South, Stillwater, OK 74078

A crossmapping technique is introduced for visualizing multiple and overlapping relations among entity types in collections of journal articles. Groups of entities from two entity types are crossplotted to show correspondence of relations. For example, author collaboration groups are plotted on the *x* axis against groups of papers (research fronts) on the *y* axis. At the intersection of each pair of author group/research front pairs a circular symbol is plotted whose size is proportional to the number of times that authors in the group appear as authors in papers in the research front. Entity groups are found by agglomerative hierarchical clustering using conventional similarity measures. Crossmaps comprise a simple technique that is particularly suited to showing overlap in relations among entity groups. Particularly useful crossmaps are: research fronts against base reference clusters, research fronts against author collaboration groups, and research fronts against term co-occurrence clusters. When exploring the knowledge domain of a collection of journal papers, it is useful to have several crossmaps of different entity pairs, complemented by research front timelines and base reference cluster timelines.

Collections of journal papers related to a scientific field are a useful source of information when mapping a knowledge domain (1). The structure within the knowledge domain is manifested in the collection of papers as groups of related entities, such as groups of papers that represent subtopics, groups of references that represent base knowledge, groups of paper authors that represent collaboration teams, groups of reference authors that represent experts, groups of journals that represent subtopic libraries, and groups of terms that represent specialized vocabularies within the knowledge domain. Exploration and visualization of these groups and the complex relations among them provides information that can be used to gain a broad and detailed understanding of the underlying knowledge domain.

Inherently, entity groups within collections of journal papers exhibit considerable "core and scatter" in group membership (2), with each group usually possessing a small core group of strongly related member entities and a much larger group of weakly related scatter members. Furthermore, there is considerable overlap in membership of entities in groups. For a thorough understanding of the structure of a knowledge domain, it is useful to visualize and understand the extent of overlap among groups in a collection of journal papers.

This article introduces a simple technique for visualizing the relations among collections of entity groups. The technique, which uses a crossmap format to show the magnitude of correspondences between all pairs of groups drawn from two differing entity types, allows visualization of relations between groups and additionally permits visualization of overlap in group membership. Using this technique, it is possible to visualize and understand the set of complex relations among the different groups that are manifested in a knowledge domain. For example, given a collection of research fronts, i.e., groups of papers reporting on the same subtopic, for each research front it is possible to identify the groups of important references, contributing author collaboration teams, groups of experts (important reference authors), and key journals.

This article is organized as follows. The reader is introduced to a simple entity-relationship model of collections of journal papers. This is followed by a discussion of important entities used for mapping knowledge domains and a discussion of co-occurrence relations that are used to cluster entities into groups. A detailed discussion of important entity groups appears, including an explanation of core and scatter and overlap of group membership. Then, the use of correspondence metrics between entity groups is discussed, followed by a detailed discussion of the proposed crossmapping technique. Finally, an example is presented, showing crossmaps produced from a collection of journal papers related to the subject of complex networks.

Entity Model of Journal Paper Collections

Using an entity-relationship model (3), collections of journal papers may be considered to be a collection of entities of differing entity types. Examples of entity types for journal paper collections include papers, paper authors, references, and paper journals. Borner, Chen, and Boyack (1) describe these entities as "units of analysis" and list the entity types most commonly used for mapping knowledge domains and also show applications of the analysis of each entity type. This article will expand on analysis of entities for knowledge mapping to include analysis of relations between different types of entities, thus extending the understanding of complex knowledge domains.

Within the collection of journal papers, entities are associated with each other. For example, each paper in the collection is associated with the authors who wrote it, the references it cites, the journal in which it was published, and the terms that appear in it. As presented here, these associations are always between pairs of entities of differing entity type. Entities of the same entity type are never directly associated. For example, papers and references are considered distinct entity types, even though references often correspond to actual papers. This separation into two distinct entity types is necessary, because papers and references represent differing concepts. A paper represents a research report, whereas a reference represents a symbol of knowledge (4). Similar considerations require designation of paper journals, reference journals, paper authors, and reference authors as separate entity types. Fig. 1 shows an ontology diagram of the entities within a collection of journal papers and the types of associations among those entities. Fig. 2 explains the symbols used in Fig. 1.

Co-Occurrence Among Entities

Co-occurrence relations among entities of the same entity type occur when two entities of the same type are associated with an

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

*To whom correspondence should be addressed. E-mail: samorri@okstate.edu.

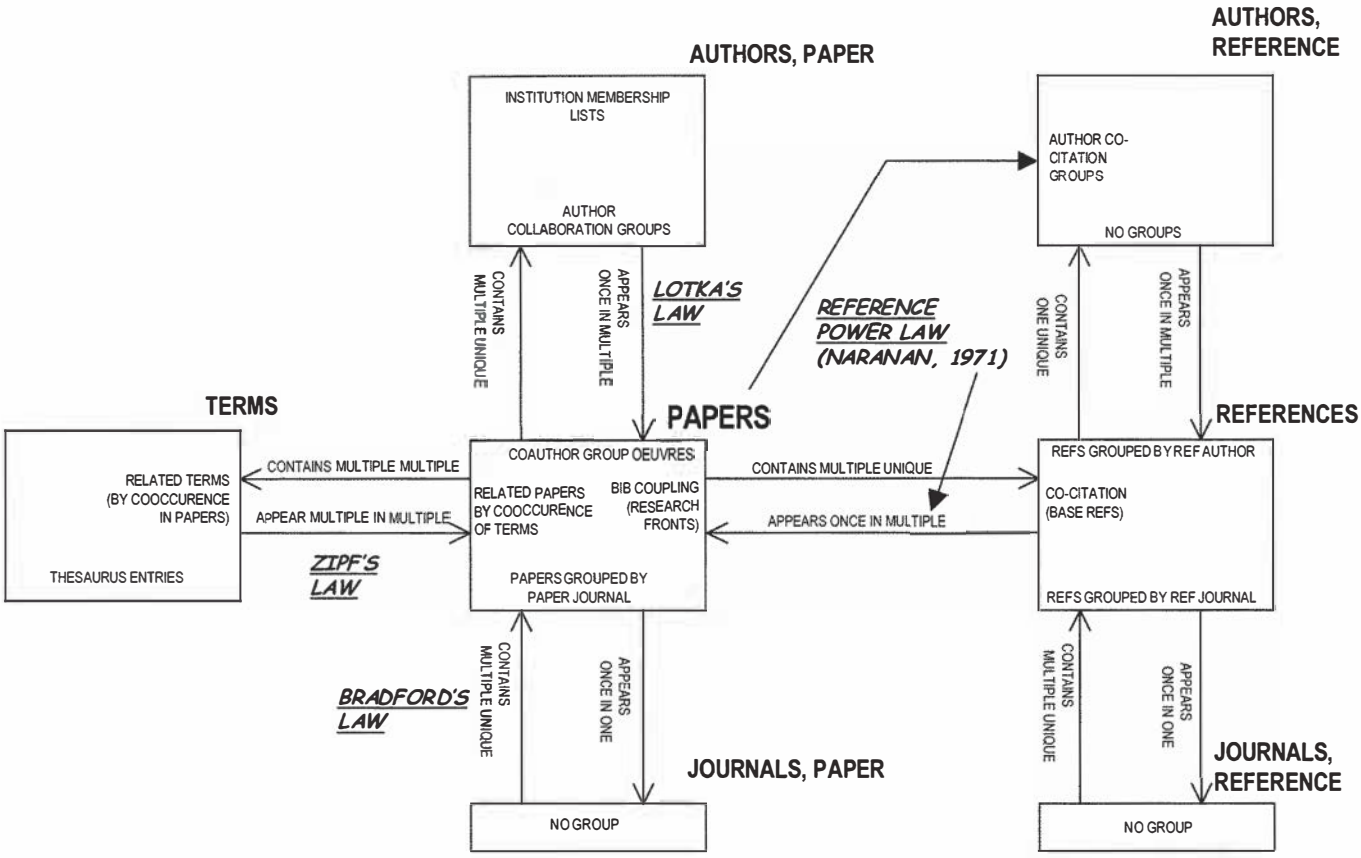


Fig. 1. Ontology diagram of the entities within a collection of journal papers, their direct relations to each other, co-occurrence groups, and core and scatter relations among those entities.

entity of a differing entity type. For example, two authors are related when they coauthor a paper, two papers are related when both cite the same reference, or two references are related when they are cited together in the same paper. Co-occurrence relations between pairs of entities often imply some meaningful relation between those entities. For example, coauthorship of papers implies that pairs of authors are collaborators, common references between papers implies that pairs of papers deal with the same research topic, and cocitation of references implies that two references are symbols of similar base knowledge.

Several of the co-occurrence relations that occur within collections of journal papers have been named and extensively studied. These relations are noted in Fig. 1 and include:

- Bibliographic coupling: Relation of pairs of papers by common references, implying a common research topic between the papers (5).
- Cocitation: Relation of pairs of references by their co-occurrence in papers, implying that those two references are symbols of similar base knowledge (6).
- Author cocitation: Relation of pairs of reference authors by their co-occurrence in papers, implying that the two authors are symbols of the same base knowledge (7).
- Coauthorship: Relation of pairs of paper authors by coauthorship of papers, implying that the two authors are members of the same collaboration team (8).

Many other co-occurrence relations are possible, as noted in Fig. 1. For example, pairs of papers related by common terms may imply a common research topic, or pairs of journals that contain papers that cite common reference authors may imply that those two journals publish papers that have a common research topic.

Entity Groups

Using similarity metrics derived from co-occurrence counts between pairs of entities, and applying clustering techniques, groups of entities possessing commonalities can be identified in the collection of journal papers. Examples of commonly studied groups of entities are noted in Fig. 1 and include (2):

- Research fronts: Groups of papers that share a common research topic (9). Derived from co-occurrence of references in papers, these groups can be considered as representing Kuhnian puzzles within a scientific field (10, 11).
- Base reference groups: Groups of references that serve as symbols of similar base knowledge (12). Derived from co-occurrence of references in papers, these groups can be considered as representing Kuhnian exemplars or paradigms.
- Reference author groups: Groups of reference authors that serve as symbols of similar base knowledge (13). They are derived from co-occurrence of reference authors in papers. Similar to base reference groups, these groups can also be considered as representing Kuhnian exemplars and paradigms, but on a more abstract scale. Reference author groups can also be considered as groups of experts (14).
- Collaboration teams: Groups of paper authors that work together. Derived from coauthorship of papers by paper authors, these groups can be considered as representing "invisible colleges" within a field (15, 16).
- Vocabularies: Groups of keyword terms. Derived from co-occurrence of terms in papers, these groups can be

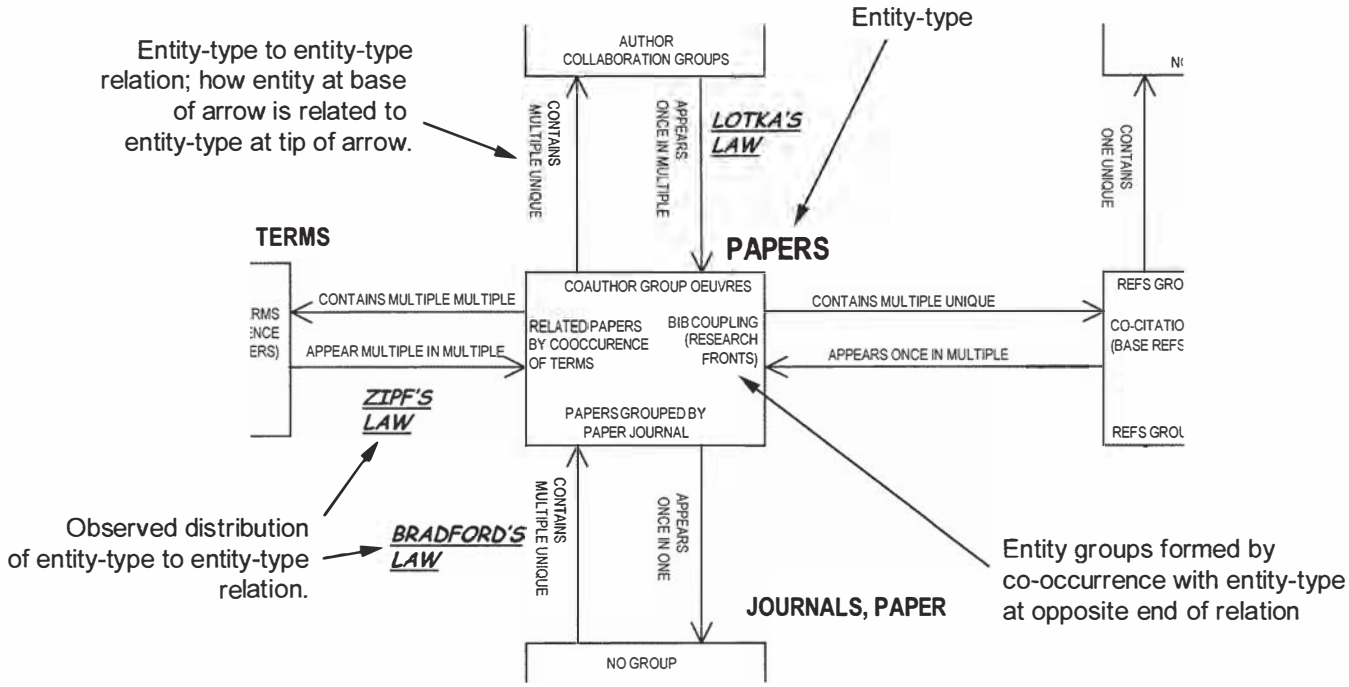


Fig. 2. Key explaining the notation in the ontology diagram of Fig. 1.

considered to represent specialized vocabularies within a research field (17).

Core and Scatter and Overlap of Group Membership

Entity groups within collections of papers exhibit core and scatter. Groups tend to possess a small set of core members that are strongly related to each other and a large number of scatter members that are weakly related (2). Furthermore, weakly related member entities are ambiguously related to many groups simultaneously. There is extensive overlap in group membership, leading to great difficulties when visualizing the knowledge domain represented by a collection of journal papers.

The core and scatter relations among entities in collections of journal papers manifest themselves as power-law distributions of entity frequency. Some of these relations, noted in Fig. 1, have been extensively researched. Example power-law relations are:

- Lotka's Law for author frequency (18).
- Bradford's Law for paper journal frequency (19).
- Zipf's Law for frequency of terms (20).
- Reference power law for frequency of references (21).

Standard clustering techniques, such as hierarchical agglomerative clustering, and standard visualization techniques, such as multidimensional scaling, do not effectively reveal the overlap of entity group membership in collections of journal papers. The crossmapping technique proposed here is designed to reveal this overlap in a field's knowledge structure by showing overlap in correspondence among groups taken from differing entity types.

Correspondence Between Groups of Entities from Differing Entity Types

Define a correspondence metric to measure the relation between a pair of entity groups drawn from differing entity types. As an example, a possible correspondence metric between a research front (a group of papers) and a base reference group is the percentage of references in the base reference group that is cited by papers in the research front. Given two collections of groups, each collection drawn from a different entity type, it is possible to build a matrix of the correspondences that exist from each group of the first entity type to each group of the second entity type.

Entity groups overlap in correspondence between groups from different entity types, e.g., a base reference cluster may have correspondence to several research fronts, or an author collaboration group may have correspondence to many term co-occurrence clusters. Knowledge of the correspondence between groups drawn from different entity types is helpful for mapping the knowledge domain associated with the collection of journal papers. Furthermore, visualization of overlapping group-to-group correspondence helps sort out complex relations among research topics, base reference groups, and research teams. In collections of journal papers, we propose several correspondence relations between groups of different entity types that are useful for knowledge mapping:

- Relation of research fronts to base reference groups. This shows what base knowledge supports specific research topics.
- Relation of research fronts to author collaboration groups. This shows what research teams work on specific research topics.
- Relation of research fronts to term co-occurrence groups. This shows what concepts are associated with specific research topics and can be helpful for labeling research fronts.
- Relation of research fronts to paper journal groups. This shows the core journals that publish papers pertinent to specific research topics.

The crossmapping technique presented here is used to visualize and explore relations between groups of entities. The technique is especially suitable to the visualization of overlap in such relations, and as such, allows the investigation of a knowledge domain through various manifestations: research fronts, base reference groups, invisible colleges, technical vocabularies, and core journals.

Description of Crossmap Visualization

The crossmapping technique presented here visualizes the matrix of correspondence magnitudes between groups from two

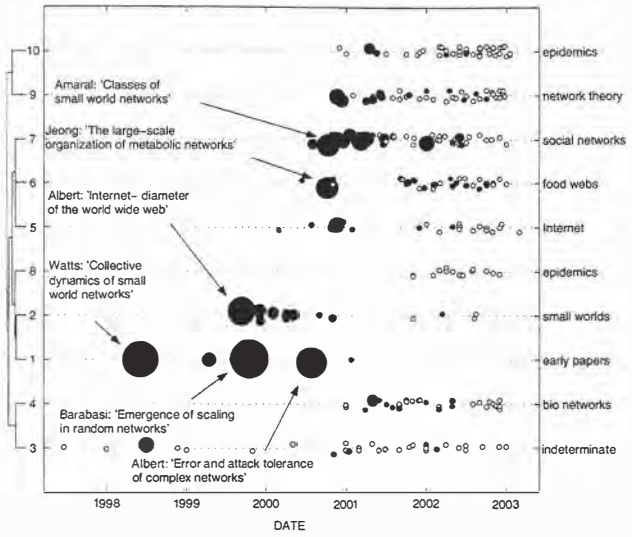


Fig. 3. Timeline of research fronts for complex networks papers. Papers are shown as circles whose size is proportional to total citations received. Filled circles are papers that have received eight or more citations in the last 12 months.

different entity types. Assuming, for example, groups of papers (entity type 1), and groups of references (entity type 2), one measure of correspondence is the number of references in a group of references that appears in a group of papers. Given N_1 groups of entity type 1 and N_2 groups of entity type 2, a N_1-by-N_2 matrix lists all of the correspondences between entity type 1 groups and entity type 2 groups. A crossmap is a visual representation of that correspondence matrix. The crossmap method is similar to MATRIX BROWSER (22), used for visualizing computer networks; DOCCUBE (23), which uses 3D matrix visualizations to aid query searches of large document collections; and GRIDL (24), an interactive system for visualizing hierarchically organized databases and library search results. The crossmap technique complements timeline visualization (11), allowing a thorough exploration of the static relations and temporal events in a collection of research fronts.

Construction of Crossmaps

To start, clustering is performed on each entity type, grouping entities according to some similarity metric. Clusters from the first entity type are mapped as rows, and clusters of the other entity type are mapped as columns. Dendrograms are added to the crossmap to show the structure of clusters being displayed. For every group at row i from entity type 1, and every group at column j from entity type 2, a circle is placed at row i and column j whose size is proportional to the magnitude of the correspondence between those two groups. Group labels are placed at row and column positions to the left and bottom of the map.

Example: A Collection of Complex Networks Papers

A collection of papers about complex networks will illustrate the use of crossmap techniques. This collection was gathered from the Institute for Scientific Information Web of Science product (www.isinet.com/products/citation/wos) by using queries to gather papers that cite several key references in the field. All groups in this set were generated by using agglomerative hierarchical clustering with Ward's method linkage on co-occurrence metrics normalized by using the cosine formula (25). Summary statistics for the entities in this collection are:

- 323 papers in 86 journals (20% of journals contain 76% of all papers).
- 11,304 citations to 6,167 references (20% of references received 54% of all citations).
- 826 authorships of 455 authors (20% of authors accounted for 52% of all authorships).

Using bibliographic coupling (5) as a co-occurrence metric, a clustering of 10 research fronts was generated. Fig. 3 shows a timeline of the research fronts. Papers are shown as circles plotted by publication date in horizontal tracks whose vertical position corresponds to positions on the clustering dendrogram shown to the left. Circle size is proportional to the number of citations received, and circles are darkened for papers that have

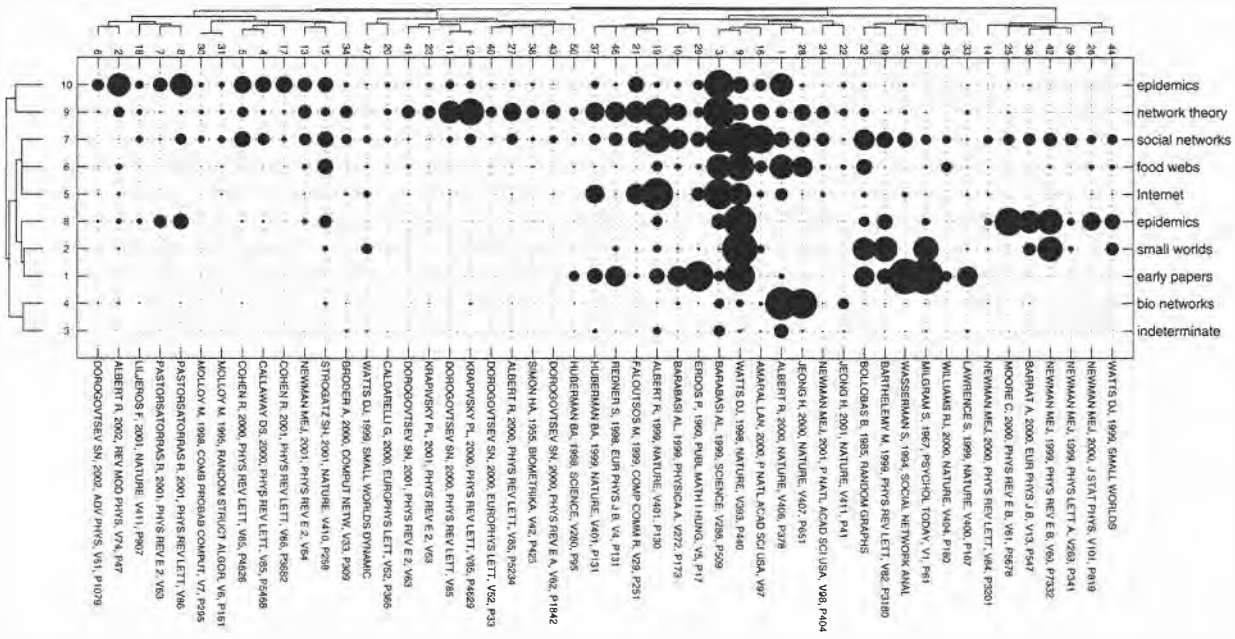


Fig. 4. Crossmap of research fronts to base reference groups for a collection of complex networks papers.

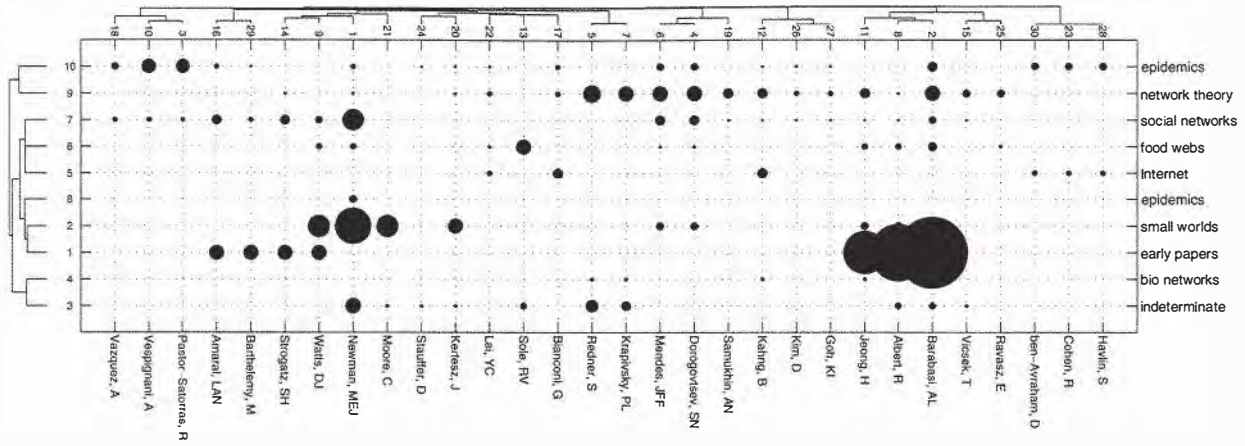


Fig. 5. Crossmap of research fronts to author collaboration groups for a collection of complex networks papers.

received eight or more citations in the last 12 months. The six most cited papers are noted in Fig. 3.

The research front labels were added after manually browsing titles of papers in each research front for themes. In future research, it may be possible to automate or semiautomate the labeling process by using correspondence to terms derived from term co-occurrence groups. When browsing titles, considerable overlap is found in themes. Note, for example, the label "epidemics" is used for both research front 8 and research front 10. Interestingly, research front 7, "social networks," has many recent papers that are currently highly cited, implying an important research topic providing current base knowledge. Research fronts 10 and 8, both labeled epidemics, are both recent, indicating an emerging topic of research.

Fig. 4 is a crossmap of research fronts to base reference clusters, which were found by using the cocitation similarity metric (6). References cited <20 times were discarded, leaving 50 references for clustering, shown individually. Correspondence was measured by counting the number of times a reference appears in papers in a research front. The dendrogram at the top of the map shows ≈8 base reference clusters. The central group, references 3, 9, and 16 in Fig. 4, are used by all research fronts except research fronts 4 and 3. Note the difference in the two epidemics research fronts: research front 10 uses references by authors Albert and Pastor-Satorras (references 1, 2, and 8), whereas research front 8 relies heavily on references by authors Moore, Barrat, and Newman (references 25, 36, and 42). It is also easy to see that research fronts 7 and 8 overlap in their use of references 14–44, but at the same time research fronts 7 and 10 overlap, using references 5–34.

Fig. 5 shows a crossmap of research fronts to author collaboration groups. Collaboration groups were found by using coauthorship counts as the similarity metric. Authors with fewer than three papers were discarded, leaving 28 authors for clustering, which are individually shown. Correspondence is measured by counting the number of times an author appears in papers in a research front. Author groups are easily discerned from the dendrogram at the top of the map. Example groups are authors 14–21 (Strogatz, Watts, Newman, and Moore), and authors 11–2, (Jeong, Albert, and Barabasi). Note the overlap of author groups 6, 2, and 18 across research fronts 10 and 7. Additionally, Albert (author 8), whose papers were used as references from research front 10, does not appear to author any papers in that research front. Other overlapping relations are evident, particularly author groups contributing to research fronts 7 and 9.

Conclusion

The crossmapping technique shown here provides an easily understood method for exploring relations in a collection of papers. In the example shown here, two types of crossmaps allow comprehension of the overlapping relations among research fronts, base reference groups, and author collaboration groups. Potential users of crossmaps are researchers exploring the literature of a scientific field to discover current research topics, base references, research teams, core journals, and the relations among them. In our experience, the method is particularly useful for mapping a domain during the initial state-of-art review phase of new research projects when the researchers are most unfamiliar with a field's literature. The method will be useful for summarizing information about a field for presentation to subject matter experts for technology forecasting. User studies need to be conducted to validate the ease of comprehension of crossmaps and identify the most useful pairs of entity types for crossmapping.

The technique has so far been applied mainly to small, well focused collections of papers (<1,000 papers.) The principal limitation to crossmapping of large collections of papers is the restricted space available on the axes for labels. The technique may be adaptable to very large collections if interactive tools are added to expand and contract levels of the clustering hierarchy.

1. Borner, K., Chen, C. & Boyack, K. W. (2002) *Annu. Rev. Inf. Sci. Technol.* **37,** 179–255.
2. White, H. D. & McCain, K. W. (1989) *Annu. Rev. Inf. Sci. Technol.* **24,** 119–186.
3. Chen, P. P. S. (1976) *ACM Trans. Database Systems* **1,** 9–36.
4. Small, H. (1978) *Social Studies Sci.* **8,** 327–340.
5. Kessler, M. M. (1963) *Am. Doc.* **14,** 10–25.
6. Small, H. G. (1973) *J. Am. Soc. Inf. Sci.* **24,** 265–269.
7. White, H. D. & Griffith, B. C. (1981) *J. Am. Soc. Inf. Sci.* **32,** 163–171.
8. Subramanyam, K. (1983) *J. Inf. Sci.* **6,** 33–38.
9. Persson, O. (1994) *J. Am. Soc. Inf. Sci. Technol.* **45,** 31–38.
10. Kuhn, T. S. (1970) *The Structure of Scientific Revolutions* (Univ. of Chicago Press, Chicago).
11. Morris, S. A., Yen, G., Wu, Z. & Asnake, B. (2003) *J. Am. Soc. Inf. Sci. Technol.* **55,** 413–422.
12. Small, H. (1997) *Scientometrics* **38,** 275–293.
13. White, H. D. & McCain, K. W. (1998) *J. Am. Soc. Inf. Sci.* **49,** 327–355.
14. Garfield, E. (1979) *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities* (Wiley, New York).
15. Kretschmer, H. (1997) *Scientometrics* **40,** 579–591.
16. Crane, D. (1972) *Invisible Colleges: Diffusion of Knowledge in Scientific Communities* (Univ. of Chicago Press, Chicago).

17. Callon, M., Courtial, J. P. & Laville, F. (1991) *Scientometrics* **22,** 155–205.
18. Lotka, A. J. (1925) *J. Wash. Acad. Sci.* **16,** 317–323.
19. Bradford, S. C. (1938) *Engineering* **137,** 85–86.
20. Zipf, G. K. (1949) *Human Behavior and the Principle of Least Effort* (Addison–Wesley, Reading, MA).
21. Naranan, S. (1971) *J. Doc.* **27,** 83–97.
22. Ziegler, E., Kunz, C., Botsch, V. & Schneeberger, J. (2002) in *Proceedings of the IEEE Sixth International Conference on Information Visualization* (IEEE, Los Alamitos, CA), pp. 361–366.
23. Mothe, J. & Chrisment, C. (2003) *J. Am. Soc. Inf. Sci. Technol.* **54,** 650–659.
24. Shneiderman, B., Feldman, D., Rose, A. & Grau, F. G. (2000) in *Fifth ACM Conference on Digital Libraries* (Association for Computing Machinery, New York), pp. 57–66.
25. Salton, G. (1989) *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer* (Addison–Wesley, Reading, MA).

User-controlled mapping of significant literatures

Howard D. White*, Xia Lin, Jan W. Buzydlowski, and Chaomei Chen

College of Information Science and Technology, Drexel University, Philadelphia, PA 19104

We apply a version of our web-based literature-mapping system to PNAS for 1971–2002, as indexed by the National Library of Medicine and the Institute for Scientific Information. Given a single input term from a user, a medical subject heading, a cocited author, or a cocited journal, PNASLINK rapidly displays views in which that term and the other 24 terms that most frequently co-occur with it in a bibliographic database are interrelated in ways suggesting fruitful combinations for document retrieval. The interrelationships are produced by two algorithms, pathfinder networks and Kohonen-style self-organizing maps. PNASLINK displays are themselves interactive interfaces that can retrieve documents from digital libraries (e.g., PNAS Online). This style of visualizing knowledge domains is called "localized" because it does not attempt to map the indexing of literatures in full but concentrates on the top terms in an "associative thesaurus" reflecting user interests. It also permits swift remappings, as the user recognizes terms worth pursuing. PNASLINK is illustrated with maps drawn from the literature of population genetics. Some comparative and evaluative comments are added, one from a domain expert indicating that the face validity of the system may be tempered by insufficient specificity in the indexing terms being mapped.

We here present two ways of rapidly mapping literatures in terms of selected indexing vocabularies. Both ways are responsive to users, and either can serve as an interface for retrieval of documents from digital libraries. Either can also complement a work that focuses on the structure of a literature, such as a research review (1). Our data are the contents of PNAS for 1971–2002, as described by medical subject headings from the National Library of Medicine (NLM) and by citation indexing from the Institute for Scientific Information (ISI).† Indexing by these organizations typifies the bibliographic control that is extended only to significant, that is, highly valued, literatures. Our software, called PNASLINK (available at http://project.cis.drexel.edu/pnas), is designed to amplify such control, by enabling customized browsing on the basis of user input.

Both of our mapping techniques exploit co-occurrences of terms in NLM and ISI bibliographic records. The terms are systematically paired, and their co-occurrences are counted in matrices. Because people can easily assimilate numeric matrices only when they are recast as pictures of some kind (2), one of our techniques transforms the counts into Kohonen-style self-organizing maps (SOMs) (3), and the other transforms them into pathfinder networks (PFNETs) (4). SOMs show frequently co-occurring terms as nodes that are spatially close. PFNETs show them as nodes with explicit ties. The two kinds of maps will be exemplified here with medical subject headings (MeSH) and cocited authors in a specialty of genetics.

Other researchers have visualized bibliographic data with PFNETs and SOMs (2, 5), but we use them to map significant literatures in real time with retrieval capabilities built into the maps (refs. 6 and 7 and cf. ref. 8). The data are initially processed by our NOAH indexing engine, a specialized database application we designed for fast computations with verbal co-occurrence data (9). With NOAH, mapping time is determined by the size of the indexing vocabulary, not by the number of documents in the database. In CONCEPTLINK, a predecessor of PNASLINK, for example, we can almost instantly create maps of MeSH terms from >12 million MEDLINE records. (CONCEPTLINK maps the co-occurring MeSH indexing of the journals in NLM's PubMed. It is available at http://project.cis.drexel.edu/conceptlink.) Once the data are indexed, the user can map and manipulate them through a unified web interface.

The maps (Figs. 1 and 2) are based on term counts solely from PNAS records because we were set that task as participants in this colloquium. Elsewhere, we have mapped terms drawn from the NLM and ISI databases in full.‡ However, even by itself PNAS is a major interdisciplinary resource, and we can easily imagine PNAS online or other journal-specific web sites offering domain visualizations like ours for the benefit of users.

In refs. 6 and 7, we discussed AUTHORLINK (http://project.cis.drexel.edu/authorlink), the version of our software that is used to map cocited authors from ISI's Arts and Humanities Citation Index for 1988–1997. The present article attests to the quick adaptability of the CONCEPTLINK/AUTHORLINK software to the authors, journals, and MeSH terms in the PNAS data. This in turn prompts us to offer a rationale for our general approach, the localized mapping of association thesauri, along with different accounts of the two main algorithms. We describe certain interactive features of our system and conclude with some fresh comparative and evaluative data, including an expert's commentary on PNASLINK as applied to the domain of population genetics.

Localized vs. Global Mapping

In their extensive review, Börner *et al.* (5) emphasize that "painting a big picture" is a main goal in domain mapping. This may lead to a strategy of mapping very large co-occurrence matrices in their entirety. Indeed, system designers have made many significant developments in software for such global portrayals of literatures, e.g., THEMESCAPE and VXINSIGHT render literatures as landscapes; GALAXIES and STARRYNIGHT render them as astral bodies (10–12). Ours, however, is an alternative way of visualizing knowledge domains, the localized mapping. Perhaps the chief difference is that the localized approach relinquishes scope to increase the user's control of the mapping process.

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

Abbreviations: NLM, National Library of Medicine; ISI, Institute for Scientific Information; SOM, self-organizing map; PFNET, pathfinder network; MeSH, medical subject headings.

*To whom correspondence should be addressed. E-mail: whitehd@drexel.edu.

†These data are extracted from *Science Citation Index Expanded* [Institute for Scientific Information, Inc. (ISI), Philadelphia, PA; Copyright ISI]. All rights reserved. No portion of these data may be reproduced or transmitted in any form or by any means without the prior written permission of ISI.

‡For restricted access to mapping of ISI's full databases, contact X.L. at xlin@drexel.edu or H.D.W.

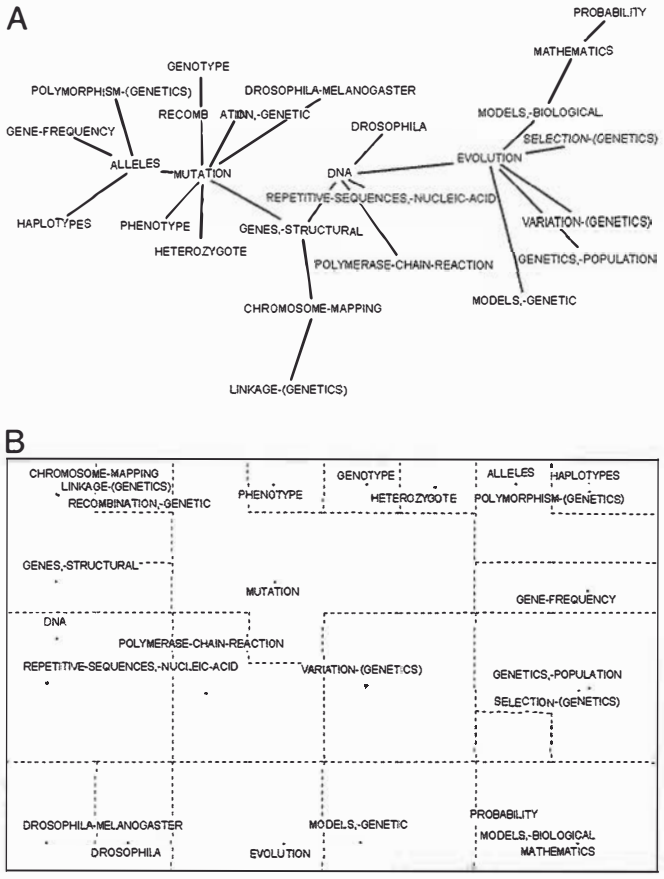


Fig. 1. (*A*) PFNET of Gene Frequency. (*B*) SOM of Gene Frequency.

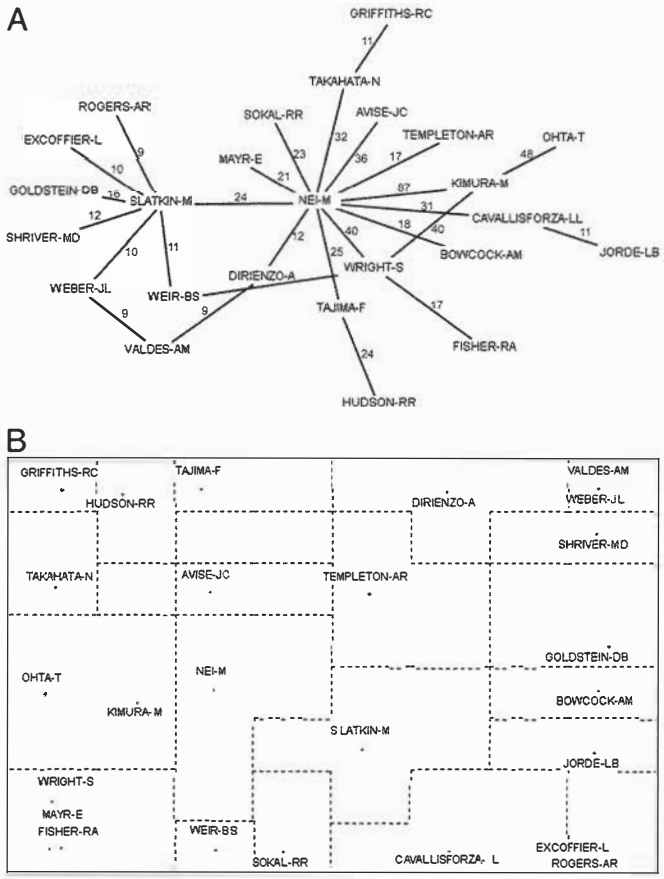


Fig. 2. (*A*) PFNET of Montgomery Slatkin. (*B*) SOM of Montgomery Slatkin.

Table 1 helps to sharpen this comparison. In global mapping, system designers present the user with a preformed view, often in 3D, of some sizeable literature. Within the panel of visualization, landscapes invite flyovers; star-fields or other constructs invite flythroughs. In the former, peaks representing major accretions of documents on some subject are likely to exert a powerful pull on the user; in the latter, document points coded as important, e.g., by differences in shape, size, or color, exert a similar pull. Essentially, the user is engaged in old-fashioned browsing, as of book titles in library stacks, but system designers may minimize or even eliminate labeling of objects in the map because labels clutter precious screen space and block the metaphorical presentation (see examples in ref. 12). The user explores the view by "visiting" or "homing in on" objects of interest, rather as in video games, but typically cannot remap the literature in pursuit of some new interest because a new map takes hours of computer time to create.

In contrast, our localized system of mapping more closely resembles online searching. The user starts the process by entering a single term at a web interface. This is consistent with the way most people search the web (13) and is intended to minimize cognitive demands on users. It is true that PNASLINK must be entered with MeSH or ISI-style terms instead of whatever word pops into the user's head, but our system includes guides that help one make the proper entries. The system responds to the entry (or "seed") term by forming a list of the terms that co-occur with it, ranked high to low by frequency. The seed term and its 24 next-highest neighbors are then exhibited as a PFNET or a SOM, which the user can switch between. Each of the two modes of mapping in PNASLINK yields different insights into the relations of the indexing terms. Both modes place the user's seed term in the locale of a limited number of other terms that are guaranteed to co-occur with it, thus customizing browsing. Any of these terms, if selected by the user, will be automatically "ANDed" with the seed term in retrieving documents.

Mapping only 25 terms at a time is an arbitrary design decision with several advantages. It allows PNASLINK to make maps on the fly in seconds. It affords the node labels, the indexing terms, enough room that they have little or no overlap, thus making them and their interrelationships the primary features of the display. It gives the user a rich, but not overwhelming, array of associations to work with. Finally, because of its speed, it permits users to create new maps on the basis of single or combined terms from an old map. Thus, instead of visiting different places in a global visualization, one moves locally from interest to interest by point-and-click remapping (which accords with Hearst's point in ref. 10 that an interactive system should let users change their search strategies as their goals change). One

Table 1. Contrasting styles of literature mapping

Global	Localized
Designer initiated	User initiated
Full matrix mapped	Small subset of matrix mapped
Map created before inquiry	Map created by inquiry
Creation time: hours	Creation time: seconds
Labeling deemphasized	Labeling emphasized
Explore by relocating on map	Explore by generating new map
Surface objects not manipulable	Surface objects manipulable
User is visitor	User is wielder

also moves by recognizing terms of interest rather than by having to guess them or look them up in a thesaurus.

The Associative Thesaurus

If the indexing terms used in the mapping are indeed controlled by a formal thesaurus, our SOMs and PFNETs provide an alternative: they display the top listings in what is sometimes called a term's associative thesaurus (2). Formal thesauri are published in hard and soft copy; associative thesauri are created ad hoc within search software. A formal thesaurus, such as NLM's *MeSH*, brings out a term's standard linguistic features, e.g., its definition, synonyms, hypernyms, and hyponyms. In contrast, an associative thesaurus shows what terms co-occur with it when it has been used to index actual publications (cf. ref. 14).

In NLM's *MeSH,* for example, the term Anthrax is related to *Bacillus anthracis* and subordinated to Bacterial Infections and Mycoses. But if Anthrax is mapped in our system, which covers the biomedical literature through 2002, its top co-occurring terms include Postal Service, a connection obviously never to be part of its entry in *MeSH.* Associative thesauri are shaped by historical contingencies, by what is being written about. That is why they may be useful for online retrieval in ways that formal thesauri are not.

Not all indexing uses subject headings and formal thesauri, of course. ISI's indexing, for example, allows searchers to retrieve the items that cite a given author. From that capability, online searchers with the right software can move to retrieving items that cite pairs of authors jointly. To people literate in a domain, frequently cocited authors may suggest nuances of meaning that are absent in standard subject indexing (for example, articles that cite both Derek de Solla Price and Diana Crane may bear on "invisible colleges" in science even if that phrase does not appear in their bibliographic records). A map of cocited authors is, in effect, an associative thesaurus of authors linked by conjoint use of their works. Again, these linkages may permit useful retrievals that are not otherwise possible (1).

Additional Capabilities

PNASLINK can produce maps not only of associated MeSH terms and cocited authors but also of cocited journals. That is, if a user supplies the name of a seed journal, such as *Gut* or *Cell*, PNASLINK maps the top 24 journals cocited with it in PNAS. Journal maps are most likely to be of interest to professional literature managers, such as serials librarians, whereas maps of MeSH and cocited authors are intended more for users in general.

Guided by our emphasis on user control, we have implemented several interactive functions for PNASLINK. For example, the system lets the user regenerate the maps after removing some terms. This is helpful, for example, in journal mapping, when one may want to eliminate omnibus journals like *Science* and *Nature* from a map to focus on more specialized titles.

PNASLINK also has alternate data models to show term relationships from different perspectives. By default, the seed term is used to generate 24 other terms, but then the counts for these pairs are obtained without reference to their counts with the seed term. However, if the user chooses the "tri-citation" option, the seed term is always required to be present with other two, and the maps are accordingly different.

Throughout the interaction process, the user can directly retrieve documents by subject through PNAS Online. Every time the user clicks on a MeSH term, it is added to a query list. When the user clicks on the find button, a separate window opens to show the documents retrieved from PNAS Online by the terms in the query list. The maps are thus a "live" interface that allows the user to interact with terms to see what documents they yield. (PNAS Online lacks ISI-type indexing, which prevents the cocited author retrieval possible in, e.g., our AUTHORLINK system).

Two Modes of Mapping

PFNETs and SOMs are dimension-reduction techniques that have been used to visualize the structure of literatures for more than a decade. In the context of the movement joining bibliometrics with document retrieval (2, 5, 10), PFNETs have been described by Fowler and colleagues (15–17), McGreevy (18), and Chen (19, 20). Analogous accounts of SOMs have been done by Lin *et al.* (21), Roussinov and Chen (22), and Chen *et al.* (23).

PFNETs. Characterizing PFNETs, Börner *et al.* (5) write, "Pathfinder algorithms take estimates of the proximities between pairs of items as input and define a network representation of the items that preserves only the most important links." Our input is pairs of terms, and the pairs are linked as output only if their co-occurrence counts are the highest (or tied-highest) in their respective vectors. By emphasizing only the most prominent links, PFNETs reduce the user's cognitive load in interpreting the most important relationships depicted in the map. These relationships in particular are highlighted as potentially fruitful for retrievals.

PFNETs were developed to portray the results of studies in which subjects' judgments of the closest semantic items were represented by the lowest weights. That is, the algorithm selects the lowest-weight (also called minimum-distance or minimum-cost) paths to render the most salient ties. However, in our matrices the closest connections are signaled by the highest co-occurrence counts. The counts therefore require a transformation (subtraction from a constant) to convert them to a distance measure before PFNETs are actually plotted.

In PFNETs, nodes represent terms, and the importance of links between them is measured by path weights, computed from term co-occurrence counts. The PFNET algorithm compares these weights over both direct (one link) and indirect (multilink) paths between nodes. It retains just those links that constitute minimum-weight paths. Such paths are required not to violate the triangle inequality $d(a,c) \leq d(a,b) + d(b,c)$, where d is the distance between points a, b, and c. These paths will be direct unless an indirect path is computed to be shorter.

The number of links in a PFNET is controlled by two parameters, r and q. These are set in our software so as to produce the sparsest possible network, which occurs when r equals infinity and q equals $n - 1$, where n is the number of nodes in the matrix.

The parameter r, which determines how path weights are computed, is lucidly explained by Fowler *et al.* (17): "Path weight, r, is computed according to the Minkowski r-metric. It is the rth root of the sum of each distance raised to the rth power for all links in a path between two nodes. Although the r-metric is continuously variable, simple interpretations exist only for $r = 1$ (path weight is the sum of the link weights in the path), $r = 2$ (path weight is the Euclidean distance), and $r =$ infinity (path weight equals the maximum link weight in the path). One advantage of $r =$ infinity is that one need only assume that the original distance estimates have ordinal properties. Another advantage is that the link structure will be preserved for any monotonic transformation of the data."[§]

The parameter q sets the range within which all paths of length q will be examined in the test of the triangle inequality (24) and removed if they violate it. The larger the value of q, the more extensive the triangle inequality constraint; therefore, links are more likely on a path that violates the rule. If q is one less than the number of nodes, then all of the potential violators are under scrutiny.

The settings $r =$ infinity and $q = n - 1$ are widely used in pathfinder research because they tend to produce networks that

[§]Quoted with permission from ref. 17.

Table 2. Top terms associated with gene frequency in two databases

Common to PNAS and PubMed	Unique to PNAS	Unique to PubMed
Alleles	*Drosophila*	Apolipoproteins E
Chromosome mapping	*Drosophila melanogaster*	Caucasoid race
DNA	Evolution	Ethnic groups
Gene frequency	Genes, structural	Genes, MHC class II
Genetics, population	Genotype	Genetic markers
Haplotypes	Heterozygote	HLA antigens
Models, genetic	Linkage, genetics	HLA-DQ antigens
Mutation	Mathematics	HLA-DR antigens
Polymerase chain reaction	Models, biological	Microsatellite repeats
Polymorphism (genetics)	Phenotype	Minisatellite repeats
Repetitive sequences, nucleic acid	Probability	Mongoloid race
Selection (genetics)	Recombination, genetic	Tandem repeat sequences
Variation (genetics)		

are highly intelligible simplifications of the data. An algorithm called a spring embedder (25) is used to enhance the maps by minimizing unsightly features such as crossed links and overlapping nodes. The finished map is virtually instantaneous once a seed term is entered.

SOMs. Unlike PFNETS, which explicitly join highly related terms, SOMs render semantic relationships through a distance metaphor. The more frequently co-occurring terms, which presumably have greater mutual relevance, occupy more proximate regions on the map. SOMs are designed to render not just the highest co-occurrence counts between terms, but rather relatively high co-occurrences across groups of terms. They are a softer-focus kind of mapping than PFNETs, but they, too, suggest specific combinations of terms on which the user might want to base retrievals.

The PNASLINK algorithm extracts the proximity relations of data in 25 dimensions, one for each of the input terms paired with all others, and seeks to preserve them as closely as possible in 2D. This process of self-organization (also known as unsupervised learning) runs over many iterative cycles. In each iteration, the images of term pairs that are strongly related in the high-dimensional space will be moved closer on the lower-dimensional space until stability is reached.

More specifically, the 2D grid of PNASLINK consists of 64 output nodes distributed in an 8-by-8 pattern. Each output node corresponds to a vector of 25 weights that are initially set as small random numbers. Each is also connected to 25 input nodes, and the latter correspond to vectors in the 25-by-25 matrix comprising all possible pairs of a seed term and the 24 terms most frequently co-occurring with it. [There are $25(24)/2 = 300$ unique pairs in the matrix, and the main diagonal, consisting of terms paired with themselves, is not used.] This co-occurrence matrix is used to train the SOM.

The account of PNASLINK's parent AUTHORLINK (6) describes the iterative training process as follows. A row from the co-occurrence matrix "is randomly selected and compared to every output node to determine a winner. Weights of the winning output nodes then are updated so that the next time this input node is presented, this output node will likely be selected again as the winner. In the meantime, nodes surrounding the winning node are similarly adjusted. The number of iterations needed to train a SOM is often determined empirically (in our case, we optimize the number of training cycles to 2,500). After the training, input vectors closest in the input space will map to the same regions in the output map. The regions are delineated by areas of nodes in which the elements with the highest value on the vectors are the same."[¶] SOMs, like PFNETs, usually take only a second or two to produce.

In interpreting SOMs, points in the same area are held to be closely related. Adjacent areas reflect stronger relationships than nonadjacent areas. Terms in large areas are more influential than terms in small areas.

Examples from Population Genetics

Fig. 1*A* is a PFNET, and Fig. 1*B* is a SOM formed with the MeSH term Gene Frequency as the seed. The result is a complex, yet still radically simplified, picture of term relations in population genetics as that subject has developed in PNAS. Fig. 2 repeats the same map types with a cocited author as seed, in this case, the population geneticist Montgomery Slatkin (University of California, Berkeley), a leading researcher in the study of gene frequencies and genetic drift. In Fig. 2*A*, the author cocitation counts have been toggled on so that they appear above the links, an option not exercised with the term co-occurrence counts in Fig. 1*A*.

The two map types suggest specific terms from the literature that can be used in document retrieval. The interface of which the maps are part has been cropped away to focus on terms that are related in ways that the literature searcher often does not know in advance. Someone interested in exploring the connection between, say, Gene Frequency and Mathematics or between, say, Slatkin and Luigi Cavalli-Sforza could click on the appropriate labels and retrieve documents in which those particular conjunctions occurred. They would be documents for which Gene Frequency and Mathematics co-occur as subject headings or in which Slatkin is cocited with Cavalli-Sforza. (Further terms may be added at will.)

In Fig. 1*A* the main nodes in the PFNET are (from left) Alleles, Mutation, Genes (Structural), DNA, and Evolution, a transition from relatively specific to relatively general terms as one moves rightward. The seed term Gene Frequency is seen to be an offshoot of the literature on Alleles. Indeed, if Gene Frequency is required to be present as a third term in all pairings in the map (the tri-citation option mentioned above), the new map has Alleles at the center with 19 of the other terms radiating directly from it.

In the SOM in Fig. 1*B*, the most central term, the one whose region touches most others, is Mutation. Gene Frequency is placed near the same terms it appeared with in the PFNET, and other connections between the PFNET and the SOM can be traced, but the SOM emphasizes different relations than the PFNET. For example, the two terms for fruit flies appear apart

[¶]Quoted from ref. 6, Copyright 2003, with permission from Elsevier.

in the PFNET, whereas the SOM brings them together at lower left.

Because Fig. 1 shows term relationships solely within PNAS, the question arises whether a mapping of Gene Frequency would differ markedly across all of the journals covered by NLM's PubMed. The latter mapping is possible through our system CONCEPTLINK. It turns out that the two maps have 13 terms in common, which demonstrates the breadth of PNAS in representing topics in genetics. (However, the PFNETs have only four links in common.) Table 2 shows the common and the unique terms. Those unique to PubMed seem more specific and more oriented toward human genetics.

Many of the MeSH terms associated with Gene Frequency in Fig. 1 appear in chapters on population genetics in introductory genetics textbooks, and they are the sort of terms that turn up in textbook glossaries (e.g., Haplotypes, Heterozygotes). Ironically, beginners at the glossary stage may know too little to profit from maps like those in Fig. 1, whereas advanced students and experts may know too much. Asked to comment on Fig. 1 as a domain expert, Slatkin said that the terms and their groupings in the two maps were intelligible, but that the MeSH terms were at such a high level of generality (e.g., Evolution, Mutation, Mathematics) that almost any way of connecting them would make some sense. (He preferred the PFNET's tighter structure to the SOM's for this reason.) He thought only mappings based on a much more specific set of seed terms, e.g., the ecology of a particular species of African millipede, would have much value for him and his students.

This is a criticism with which many people might agree, and progress in bibliographic visualizations like ours may well lie in adding capabilities to map specific natural-language "co-words" from the titles, abstracts, or full texts of documents (8, 26, 27). Possibly the chief beneficiaries of MeSH (or other controlled-vocabulary) mapping will be neither beginners nor subject experts, but "in-between" persons, such as librarians, subject indexers, science writers, journal editors, and teachers as they browse the many research areas to which they come as outsiders.

Slatkin found his own cocited author maps readily interpretable. He was acquainted with every name that appears in Fig. 2. In the PFNET (which he again preferred), he identified the main structural feature, the clusters around himself and Masatoshi Nei, as representing two slightly different subject areas. Both the Nei group and the Slatkin group, he said, have contributed to the literature on genetic flow and population structure, but the Slatkin group has contributed relatively more to the literature on microsatellites (short, repetitive sequences of DNA). Hence, the PFNET was picking up a division he found meaningful.

Many combinations of linked names in Fig. 2*A* are coherent in the sense that they yield sensible internet retrievals. However, a stricter test for the coherence of a particular domain is whether an expert can rapidly and accurately predict why two authors are linked. Given a random pair from Fig. 2*A* (Nei and R. R. Sokal), Slatkin guessed that the link between them was caused by frequent cocitation of Sokal's book *Biometry* with Nei's works on the computation of standard genetic distance. A subsequent retrieval of the articles cociting the pair bore this out.

The SOM in Fig. 2*B* picks up some of the same dyadic structure as the PFNET, such as the connections between Ohta and Kimura, Tajima and Hudson, Takahata and Griffiths, Cavalli-Sforza and Jorde, and Valdes and Weber (which may reflect coauthorships as well as cocitation ties). Slatkin and Nei remain central figures, but are joined by Avise and Templeton. Interestingly, at the lower left the SOM conjoins Wright, Mayr, and Fisher, who represent the older, pioneering generation in statistical genetics. The SOM algorithm is able to bring this out solely on the basis of their overall cocitation profiles.

Other Reactions to the Map Types

If PFNETs seem directive about term relationships, SOMs are merely suggestive. However, their greater ambiguity is perhaps a virtue. Using AUTHORLINK, the forerunner of PNASLINK, Buzydlowski (9) found that SOMs outperformed PFNETs in capturing the mental models of 20 experts in selected fields of the humanities. These were SOMs and PFNETs devoted to cocited authors, exactly like those in Fig. 2.

The experts' mental models were elicited by having them sort cards bearing authors' names into intuitively meaningful piles. Their task was to show how they would group, first, the 24 authors most highly cocited with Plato (almost all quite famous) and, second, the 24 authors most highly cocited with an individual author of the expert's choice. The matrices of card-sort groupings were compared with matrices of the groupings produced by PFNET linkages and SOM positionings. For both the Plato trial, which all experts participated in, and the individual-author trials, which were unique to each expert, SOMs agreed with the card-sort data better than PFNETs. In the Plato trial, both SOMs and PFNETs were highly correlated with the pooled card-sort data (SOMs, $r = 0.97$; PFNETs, $r = 0.78$), but these correlations were significantly different at $P < 0.001$. In the individual-author trials, a t test of mean agreement scores favored SOMs significantly at $P < 0.01$. The experts were nevertheless about equally divided in their preferences for one map type over the other.

In other, less formal trials, we have found that some experts object when maps of either type differ from their mental models of how the subject headings or authors in their fields are connected. With respect to this criticism, it should be borne in mind that the maps are pictures of the database. They show term associations that have developed as authors and indexers actually create literatures, in the present case, solely within PNAS, and these will often differ from the terminological hierarchies one finds in individual heads, not to mention textbooks, thesauri, or other databases (compare Table 2). In fact, the maps should be taken as new information, not as "erroneous" attempts to generate preexisting hierarchies from bibliographic data. The ongoing task is to find which types of maps and which types of terms are most useful to particular clienteles.

1. White, H. D. (1990) in *Scholarly Communication and Bibliometrics*, ed. Borgman, C. (Sage, Newbury Park, CA), pp. 84–106.
2. White, H. D. & McCain, K. W. (1997) *Annu. Rev. Inf. Sci. Technol.* **32,** 99–168.
3. Kohonen, T. (1997) *Self-Organizing Maps* (Springer, New York), 2nd Ed.
4. Schvaneveldt, R. W., ed. (1990) *Pathfinder Associative Networks: Studies in Knowledge Organization* (Ablex, Norwood, NJ).
5. Börner, K., Chen, C. & Boyack, K. W. (2003) *Annu. Rev. Inf. Sci. Technol.* **37,** 179–255.
6. Lin, X., White, H. D. & Buzydlowski, J. (2003) *Inf. Process. Manag.* **39,** 689–706.
7. Buzydlowski, J. W., White, H. D. & Lin, X. (2003) in *Visual Interfaces to Digital Libraries*, Lecture Notes in Computer Science 2539, eds. Börner, K. & Chen, C. (Springer, Berlin), pp. 133–144.
8. Chen, H., Lally, A. M., Zhu, B. & Chau, M. (2003) *J. Am. Soc. Inf. Sci. Technol.* **54,** 683–694.
9. Buzydlowski, J. W. (2003) Ph.D. thesis (Drexel University, Philadelphia).
10. Hearst, M. (1999) in *Modern Information Retrieval*, eds. Baeza-Yates, R. & Ribeiro-Neto, B. (Addison–Wesley, New York), pp. 257–323.
11. Chen, C. (2003) *Mapping Scientific Frontiers: The Quest for Knowledge Visualization* (Springer, London).
12. Dodge, M. & Kitchin, R. (2001) *Atlas of Cyberspace* (Addison–Wesley, Harlow, U.K.).
13. Jansen, B. J., Spink, A. & Saracevic, T. (2000) *Inf. Process. Manag.* **36,** 207–227.
14. Schatz, B. R., Johnson, E. H., Cochrane, P. A. & Chen, H. (1996) in *Proceedings of the First ACM International Conference on Digital Libraries*, eds. Fox, E. A. & Marchionini, G. (Association for Computing Machinery, New York), pp. 126–133.
15. Fowler, R. H. & Dearholt, D. W. (1990) in *Pathfinder Associative Networks: Studies in Knowledge Organization*, ed. Schvaneveldt, R. W. (Ablex, Norwood, NJ), pp. 165–178.

16. Fowler, R. H., Fowler, W. A. L. & Wilson, B. A. (1991) in *Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, ed. Bookstein, A. (Association for Computing Machinery, New York), pp. 142–151.
17. Fowler, R. H., Wilson, B. A. & Fowler, W. A. L. (1992) *Information Navigator: An Information System Using Associative Networks for Display and Retrieval*, Technical Report NAG9-551,92-1 (Department of Computer Science, University of Texas-Pan American, Edinburg).
18. McGreevy, M. W. (1995) *A Relational Metric, Its Application to Domain Analysis, and an Example Analysis and Model of a Remote Sensing Domain*, National Aeronautics and Space Administration Technical Memorandum 119358 (Ames Research Center, Moffett Field, CA).
19. Chen, C. (1998) *J. Vis. Lang. Comput.* **9,** 267–286.
20. Chen, C. (1999) *Inf. Process. Manag.* **35,** 401–420.
21. Lin, X., Soergel, D. & Marchionini, G. (1991) in *Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Design in Information Retrieval*, ed. Bookstein, A. (Association for Computing Machinery, New York), pp. 262–269.
22. Roussinov, D. & Chen, H. (1998) *Commun. Cognit. Artificial Intelligence* **15,** 81–112.
23. Chen, H., Houston, A. L., Sewell, R. R. & Schatz, B. R. (1998) *J. Am. Soc. Inf. Sci.* **49,** 582–603.
24. Tversky, A. & Gati, I. (1982) *Psychol. Rev.* **89,** 123–154.
25. Kamada, T. & Kawai, S. (1989) *Inf. Process. Lett.* **31,** 7–15.
26. Ding, Y., Chowdhury, G. G. & Foo, S. (2000) *Inf. Process. Manag.* **37,** 817–842.
27. Ibekwe-San Juan, F. & San Juan, E. (2002) *Knowl. Org.* **29,** 181–197.

Searching for intellectual turning points: Progressive knowledge domain visualization

Chaomei Chen*

College of Information Science and Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104-2875

This article introduces a previously undescribed method progressively visualizing the evolution of a knowledge domain's cocitation network. The method first derives a sequence of cocitation networks from a series of equal-length time interval slices. These time-registered networks are merged and visualized in a panoramic view in such a way that intellectually significant articles can be identified based on their visually salient features. The method is applied to a cocitation study of the superstring field in theoretical physics. The study focuses on the search of articles that triggered two superstring revolutions. Visually salient nodes in the panoramic view are identified, and the nature of their intellectual contributions is validated by leading scientists in the field. The analysis has demonstrated that a search for intellectual turning points can be narrowed down to visually salient nodes in the visualized network. The method provides a promising way to simplify otherwise cognitively demanding tasks to a search for landmarks, pivots, and hubs.

The primary goal of knowledge domain visualization (KDViz) is to detect and monitor the evolution of a knowledge domain (1). Progressive knowledge domain visualization is specifically concerned with techniques that can be used to identify temporal patterns associated with significant contributions as a domain advances.

Many aspects of a scientific field can be represented in the form of a scientific network, such as scientific collaboration networks (2), social networks of coauthorship (3), citation networks (4), and cocitation networks (5). Scientific networks constantly change over time. Some changes are relatively moderate; some can be dramatic. Understanding the implications of such changes is essential to everyone in a scientific field.

Researchers have been persistently searching for underlying mechanisms that may explain various changes and patterns in scientific networks. On the other hand, this is an ambitious and challenging quest because of the scale, diversity, and dynamic nature of scientific networks that one has to deal with. In this article, we introduce a previously undescribed method designed to reduce some of the complexities associated with identifying key changes in a knowledge domain. We focus on cocitation networks, although we expect that the method is applicable to a wider range of networks.

The key elements of the method draw their strength from a divide-and-conquer strategy. A time interval is divided into a number of slices, and an individual cocitation network is derived from each time slice. The time series of networks are merged. Major changes between adjacent slices are highlighted in a panoramic visualization of the merged network. The primary motivation of the work is to simplify the search for significant papers in a knowledge domain's literature so that one can search for visually salient features, such as landmark nodes, hub nodes, and pivot nodes, in a visualized network. The entire progressive visualization process is streamlined and implemented in a computer system of the author called CITESPACE.

The rest of this paper is organized as follows. We first review prior studies of the growth of a knowledge domain and then identify the key issues to be addressed by our method. The progressive visualization method is described and illustrated with an example in which we identify intellectual turning points in the field of superstring in theoretical physics. Identified articles associated with visually salient features are validated with the leading scientists in the field of superstring.

Related Work

Two strands of research are relevant. The focus of our research can be expressed in two key questions. How does a scientific field grow? What has been done for visualizing temporal patterns, especially in relation to network evolution?

Scientific Revolutions

The most widely known model of science is Thomas Kuhn's *Structure of Scientific Revolutions* (6), in which science is characterized by transitions from normal science to science in crisis and from crisis to a scientific revolution. Kuhn's theory suggests that scientific revolutions are a crucial part of science. The notion of paradigm shift is widely known in virtually all scientific disciplines. Kuhn's model has generated profound interest in detecting and monitoring paradigm shifts through the study of temporal patterns in cocitation networks. Small (7) identified and monitored the changes of research focus in collagen research in terms of how clusters of most cocited articles change over consecutive years. Small's study predated many modern visualization techniques. However, the representations of cocitation clusters were isolated from one year to another; significant temporal patterns or transitions may go unnoticed if they fall between the clusters from different years.

In our earlier work (8), we used animated visualization techniques to reconstruct citation and cocitation events in their chronological order so that one can examine the growth history of a domain in a broader context in a similar way to how we play a video in a fast-forward mode. The animated visualization enabled us to identify paradigm-like clusters of cocited articles corresponding to significant changes in the field of superstring, the same topic we will revisit with our method, but the visual features of some of the groundbreaking articles were not distinct enough to lend themselves to a simple visual search. Our earlier methodology did not include time slicing, multiple thresholding, and merging. One of our main objectives is therefore to improve visualization techniques so that groundbreaking articles can be characterized by distinguishable visual features.

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

Abbreviation: KDViz, knowledge domain visualization.

*E-mail: chaomei.chen@cis.drexel.edu.

The concept of a research front is also relevant to how science grows. A research front consists of transient clusters of most recently cited works in the literature of a scientific field (9). A research front represents the state of the art of a field, and research fronts move along with the underlying scientific field as new articles replace existing articles.

The recent interest in complex network analysis is a potentially fruitful route to improve our understanding of scientific networks as well as general networks (10). Studies in complex network analysis, especially in relation to small-world and scale-free networks, focus on two broad issues, namely topological properties and generative mechanisms of networks. Various growth models such as preferential attachment (10–12) have been developed in the study of network evolution. However, much of the work has concentrated on abstract network representations rather than on concrete networks and their practical implications. We emphasize the integral role of the semantics of such networks in understanding the profound dynamics of network evolution. We expect that the progressive method described in this article can provide a useful instrument for examining the evolution of a scientific network, and that the concrete example of network evolution can lead to insights into a broader range of networks.

Visualizing Temporal Patterns

A good example of visualizing thematic changes in a collection of text documents is THEMERIVER (13), which uses the metaphor of a thematic river to depict temporal changes of word frequencies. An intensified theme can be identified if one can detect increasingly widened word frequency streams. This is a relatively straightforward task, given that one needs only to tell how much the width of a stream changes over time. In contrast, it is often much more complicated to detect temporal patterns in higher-dimensional data or higher-order relationships.

A number of methods such as INDSCAL (14), PROCRUSTES analyses (15), and thin-plate splines and deformation analysis (16) can be used to compare dimensional representations. PROCRUSTES analysis, for example, aligns two configurations by stretching and rotating operations so that the remaining differences are where the two configurations really differ. Similarly, thin-plate display renders the difference between two configurations as a deformed plate. The degree of the deformation indicates the extent to which the two configurations differ. Such methods are efficient in detecting local and short-range discrepancies between two almost identical configurations, but the performance degrades if the discrepancies are long range in nature or a substantial part of the configurations is involved. In KDViz, we need to consider both short- and long-range changes between two adjacent snapshots of a domain, although few empirical studies have examined these techniques in the context of KDViz.

A network can change over time in various ways and can change its topology by adding new nodes and links as well as removing existing nodes and existing links. A network can also change the intrinsic attributes of its nodes and links; for example, citations of articles in a cocitation network tend to increase over time.

Much of the existing approaches to visualizing the evolution of a network falls into one of two categories: the slide-show and panorama approaches. Just like in a slide show, the former aims to highlight the changes as the viewer moves forward, sometime back and forth, in a time series of snapshots. The latter aims to pack synthesized temporal changes into a single image.

The slide-show approach has several advantages, including being easy to implement and flexible to use. This approach often provides additional visual aids to help viewers identify changes between adjacent snapshots. Recent examples include the visualization of how a discourse evolves as a network of words (17) and the visualization of semantic structures across different time planes (18). However, research in perceptual cognition has shown that comparing two images back and forth can be cognitively vey demanding and prone to error.

The panorama approach aims to depict temporal as well as spatial changes in such a way that viewers can detect a trend or a pattern by studying a single image. This approach could minimize the disturbance to the viewer's mental model (19). Related work in this area includes incremental graph drawing (20) and the timed network display function in PAJEK (21). Our earlier work on using animated visualization techniques to depict temporal changes in a cocitation network also belongs to this category (8).

Progressive Visualization Issues

A progressive visualization method aims to visualize the evolution of a network over time. The following three issues need to be addressed for visualizing time-sliced networks: (*i*) Improving the clarity of individual networks; (*ii*) highlighting transitions between adjacent networks; and (*iii*) identifying potentially important nodes.

The first issue is concerned with the clarity of individual networks' representations. One of the major aesthetic criteria established by research in graph drawing is that link crossings should be avoided whenever possible. A network visualization with the least number of edge crossings is regarded as not only aesthetically pleasing but also more efficient to work with in terms of the performance of relevant perceptual tasks (22). The number of link crossings may be reduced by pruning various links in a network. Minimum spanning trees and Pathfinder network scaling are commonly used algorithms. The major advantages and disadvantages of these scaling techniques are further analyzed below.

The second issue is concerned with progressively merging two adjacent networks, so that one can identify which part of the earlier network is persistent in the new network, which part of the earlier network is no longer active in the new network, and which part of the new network is completely new. Much of the novelty of our method is associated with the way we address this issue.

The third issue is concerned with the role of visually salient features in simplifying search tasks for intellectual turning points. Visually salient nodes include landmark nodes, pivot nodes, and hub nodes.

Issue 1: Improving the Clarity of Networks

Cocitation networks often have a vast number of links, and displaying links indiscriminately is the primary cause of clutter. There are two general approaches to reduce the number of links in a display: threshold- and topology-based approaches. In the threshold-based approach, the elimination of a link is determined solely by whether the link's weight exceeds a threshold. In contrast, in a topology-based approach, the elimination of a link is determined by a more extensive consideration of intrinsic topological properties; therefore, such approaches tend to preserve certain topological intrinsic properties more reliably, although the computational complexity tends to be higher.

Pathfinder network scaling is originally developed by cognitive scientists to build procedural models based on subjective ratings (23–25). It uses a more sophisticated link-elimination mechanism compared to minimum spanning tree (MST) and can remove a large number of links and retain the most important ones. Given a network, one can derive a unique Pathfinder network that contains all of the alternative MSTs of the original network. MST is increasingly a strong candidate in a series of KDViz studies (8, 26–28).

The goal of Pathfinder network scaling, in essence, is to prune a dense network. The topology of a Pathfinder network is

determined by two parameters, r and q. The r parameter defines a metric space over a given network based on the Minkowski distance so that one can measure the length of a path connecting two nodes in the network. The Minkowski distance becomes the familiar Euclidean distance when $r = 2$. When $r = \infty$, the weight of a path is defined as the maximum weight of its component links, and the distance is known as the maximum value distance.

Given a metric space, a triangle inequality can be defined as follows,

$$w_{ij} \leq (\Sigma_k w^r n_k n_{k-1})^{1/r},$$

where w_{ij} is the weight of a direct path between i and j, $wn_k n_{k+1}$ is the weight of a path between n_k and n_{k+1}, for $k = 1, 2, \ldots, m$. In particular, $i = n_1$ and $j = n_k$. In other words, the alternative path between i and j may go all the way around through nodes $n_1, n_2, \ldots, n_k$, so long as each intermediate links belong to the network.

If w_{ij} is greater than the weight of alternative path, then the direct path between i and j violates the inequality condition. Consequently, the link *i-j* will be removed, because it is assumed that such links do not represent the most salient aspects of the association between the nodes i and j.

The q parameter specifies the maximum number of links that alternative paths can have for the triangle inequality test. The value of q can be set to any integer between 2 and $N - 1$, where N is the number of nodes in the network. If an alternative path has a lower cost than the direct path, the direct path will be removed. In this way, Pathfinder reduces the number of links from the original network, whereas all of the nodes remain untouched. The resultant network is also known as a minimum-cost network.

The strength of Pathfinder network scaling is its ability to derive more accurate local structures than other comparable algorithms, such as multidimensional scaling and minimum spanning tree. However, the Pathfinder algorithm is computationally expensive; the published algorithm is in the class of $O(N^4)$. KDViz approaches built on the Pathfinder network scaling algorithm have a potential bottleneck if one needs to deal with large networks. The maximum pruning power of Pathfinder is achievable with $q = N - 1$ and $r = \infty$; not surprisingly, this is also the most expensive one, because all of the possible paths must be examined for each link. In addition, the algorithm requires a large amount of memory to store the intermediate distance matrices. This is the first of the three issues our method is to deal with. The method follows a divide-and-conquer strategy.

Issue 2: Merging Heterogeneous Networks

The second issue identified above is concerned with progressively merging two temporally adjacent networks. Depending on the nature of a knowledge domain, networks to be merged could be heterogeneous as well as homogeneous in terms of intrinsic topological properties and additional attributes of nodes and links. For example, intellectual structures of a knowledge domain before and after a major conceptual revolution are likely to be fundamentally different as new theories and evidence become predominant. Cocitation networks of citation classics in a field are likely to differ from cocitation networks of newly published articles. The key question is, what is the most informative way to merge potentially diverse networks?

A merged network needs to capture important changes over time in a knowledge domain's cocitation structure. We need to find when and where the most influential changes took place so that the evolution of the domain can be characterized and visualized. Few studies in the literature investigated network merge from a domain-centric perspective. The central idea of our method is to visualize how different network representations of an underlying phenomenon can be informatively stitched together.

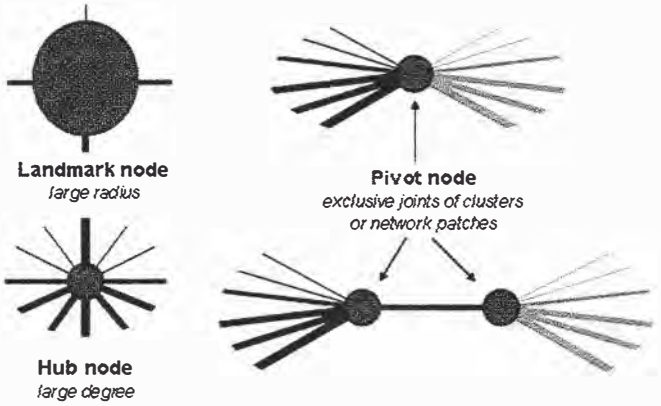


Fig. 1. Three types of visually salient nodes in a cocitation network.

Issue 3: Visually Salient Nodes in Merged Networks

The third issue addressed by our method is concerned with the identification of potentially important articles in a cocitation network. The importance of a node in a cocitation network can be quickly identified by the local topological structure of the node and by additional attributes of the node. We are particularly interested in three types of nodes: (*i*) landmark, (*ii*) hub, and (*iii*) pivot nodes (see Fig. 1).

A landmark node has extraordinary attribute values. For example, a highly cited article tends to provide an important landmark regardless of how it is cocited with other articles. Landmark nodes can be rendered by distinctive visual–spatial attributes such as size, height, or volume. A hub node has a relatively large node degree; a widely cocited article is a good candidate for significant intellectual contributions. A high-degree hub-like node is also easy to recognize in a visualized network. Both landmark and hub nodes are commonly used in network visualization. Although the concept of pivot nodes is available in various contexts, the way they are used in our method is previously undescribed. Pivot nodes are joints between different networks; they are either the common nodes shared by two networks or the gateway nodes that are connected by internetwork links. Pivot nodes have an essential role in our method.

Methods

The method includes the following procedural steps: time slicing, thresholding, modeling, pruning, merging, and mapping. Although pruning is not always necessary, it is a potentially valuable option when dealing with a dense network. All steps are implemented in CITESPACE.

Procedure. The input to CITESPACE is a set of bibliographic data files in the field-tagged Institute for Scientific Information[†] Export Format. The outputs of CITESPACE include visualized cocitation networks; each network is shown in a separate interactive window interface.

Time Slicing. The entire time interval can be sliced into equal-length segments. The length of each segment can be as short as a year or as long as the entire interval. If appropriate data

[†]These data are extracted from *Science Citation Index Expanded* [Institute for Scientific Information, Inc. (ISI), Philadelphia, PA; Copyright ISI]. All rights reserved. No portion of these data may be reproduced or transmitted in any form or by any means without the prior written permission of ISI.

become available, it is possible to slice the data thinner to make monthly or weekly segments. Currently, sliced segments are mutually exclusive, although overlapping segments could be an interesting alternative worth exploring.

Thresholding. Citation and cocitation analysis typically sample the most highly cited work, the cream of crop, with a single constant threshold. However, a single constant threshold is a crude sampling mechanism if the citation patterns over an extended period are being considered. By default, both citations and cocitations are calculated within each time slice, as opposed to across all time slices.

Time slicing provides the flexibility to tailor a threshold more closely to the characteristics of citation and cocitation activities in each individual time slice. This flexibility is expected to reduce the bias associated with a single one-size-fits-all threshold. One can even compare and merge two very different networks within this framework, for example, a network of articles from Nobel Prize-winning scientists and a network of technical reports. The key questions are: what is the common ground between two networks? How can one extract insights into the internetwork relationship from such common ground? A flexible threshold configuration can find a common ground more easily.

The cocitation network in a given time slice is determined by three thresholds: citation, cocitation, and cosine coefficient thresholds. In CITESPACE, the user needs to select desired thresholds for three specific time slices, namely the beginning, middle, and ending slices. CITESPACE automatically assigns interpolated thresholds to the remaining slices. In practice, the user starts with an arbitrary threshold configuration and then adjusts thresholds accordingly based on the reported statistics such as the citation population and the numbers of nodes and links in a network.

In the citation world, articles are not created equal. Some articles have much more than their fair share of citations, some have less, and some have none at all. Citations depend on many underlying factors. For example, success breeds success; a highly cited article is likely to receive more citations than a currently less frequently cited article. To detect intellectual turning points, we are particularly interested in articles that have rapidly growing citations. In the following superstring example, we use a simple model to normalize the citations of an article within each time slice by the logarithm of its publication age, the number of years elapsed since its publication year. The rationale is to highlight articles that increased most in the early years of publication. More sophisticated models can be derived based on citation distribution models of a given dataset and a model of the growth and decay of scientific citations (29). Building such models is significant and challenging in its own right.

Modeling. By default, cocitation counts are calculated within each time-sliced segment. Cocitation counts are normalized as cosine coefficients, $cc_{cosine}[i, j] = cc[i, j]/sqrt(c[i], c[j])$, where $cc[i, j]$ is the cocitation count between documents i and j, and $c[i]$ and $c[j]$ are their citation counts, respectively. The user can specify a selection threshold for cocitation coefficients; the default value is 0.15.

Alternative measures of cocitation strengths are available in the information science literature, such as Dice and Jaccard coefficients. In earlier studies, we used Pearson's correlation coefficients. Recently, researchers began to examine how Pearson's correlation coefficients transform the underlying structure of a cocitation network (30), but available evidence is still inconclusive (31, 32). Although the impact of various cocitation metrics on the resultant network visualizations is worth pursuing, the topic is beyond the scope of this article.

Pruning. Effective pruning can reduce link crossings and improve the clarity of the resultant network visualization. CITESPACE supports two common network-pruning algorithms, namely Pathfinder and minimum spanning tree. The user can select to prune individual networks only or the merged network only or to prune both. Pruning increases the complexity of the visualization process. In the following section, visualizations with local pruning and global pruning are presented.

In this article, we concentrate on Pathfinder-based pruning. To prune individual networks with Pathfinder, the parameters q and r were set to N_k - 1 and ∞, respectively, to ensure the most extensive pruning effect, where N_k is the size of the network in the kth time slice. For the merged network, the q parameter is (ΣN_k)- 1, for $k = 1, 2$.

Merging. The sequence of time-sliced networks is merged into a synthesized network, which contains the set union of all nodes ever to appear in any of the individual networks. Links from individual networks are merged based on either the earliest establishment rule or the latest reinforcement rule. The earliest establishment rule selects the link that has the earliest time stamp and drops subsequent links connecting the same pair of nodes, whereas the latest reinforcement rule retains the link that has the latest time stamp and eliminates earlier links.

By default, the earliest establishment rule applies. The rationale is to support the detection of the earliest moment when a connection was made in the literature. More precisely, such links mark the first time a connection becomes strong enough with respect to the chosen thresholds.

Mapping. The layout of each network, either individual time-sliced networks or the merged one, is produced by using Kamada and Kawai's algorithm (33). The size of a node is proportional to the normalized citation counts in the latest time interval. Landmark nodes can be identified by their large discs. The label size of each node is proportional to citations of the article, thus larger nodes also have larger-sized labels. The user can enlarge

Table 1. Time slicing and threshold settings for a set of small networks, where f_c is the citation frequency threshold and f_{cc} is the cocitation frequency threshold

Time slices	f_c	f_{cc}	Cite space size	Top cited	Sample, %	cc (cosine ≥0.15)
1985–1987	3	1	604	16	2.65	58
1988–1990	10	3	2,740	15	0.55	30
1991–1993	50	7	12,214	18	0.15	62
1994–1996	60	10	16,147	19	0.12	53
1997–1999	80	10	19,716	20	0.10	60
2000–2002	85	15	22,449	20	0.09	54
2003	25	10	9,594	13	0.14	34
Total (unique)			83,464	121 (82)	Mean 0.54	Total 351

cc, cocitation.

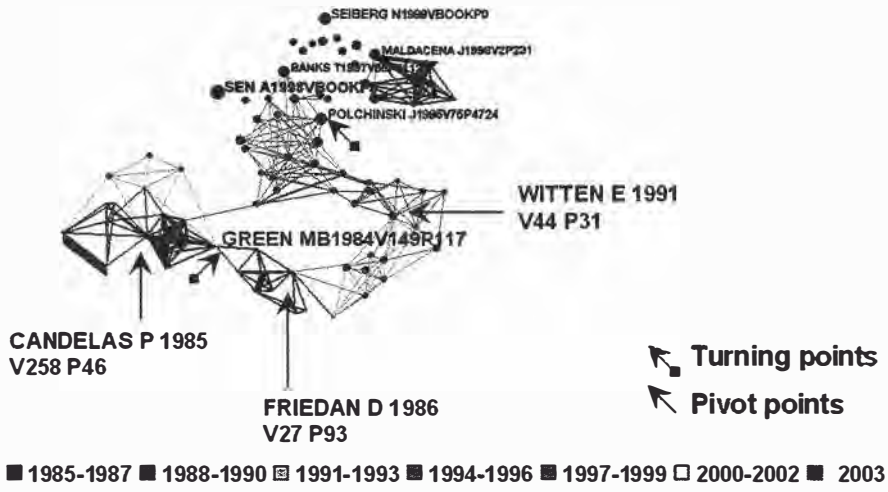


Fig. 2. An 82-node merged network without global pruning. See a color version at www.pages.drexel.edu/~cc345/citespace/Figure 2.png.

font sizes at will, and both the width and the length of a link are proportional to the corresponding cocitation coefficient. The color of a link indicates the earliest appearance time of the link with reference to chosen thresholds.

Visually salient nodes such as landmarks, hubs, and pivots are easy to detect by visual inspection. CITESPACE currently does not include any algorithms to detect such nodes computationally. Instead, the visual effect is a natural result of slicing and merging, although additional computational metrics may enhance the visual features even further. A useful computational metric should reflect the degree of a node, and it should also take into account the heterogeneity of the node's links. The more dissimilar links a node connects to others, the more likely the node has a pivotal role to play. In the following example, we consider only nodes that have a degree of 10 or higher for visual inspection.

Superstring. The method is applied to the visualization of how cocitation networks of superstring in theoretical physics evolved over time. Two superstring revolutions are documented over the last two decades: one in the mid-1980s and one in the mid-1990s (34). We reported animated visualizations of the superstring cocitation networks by using a single constant citation threshold, and Pearson's correlation coefficients were used to measure the strength of each cocitation link. The changes of the cocitation network were animated by the growing height of a citation bar and as a state transition process. Although key articles to the revolutions were identifiable in the resultant animations, they did not quite lend themselves to a simple visual inspection. We expect that the progressive visualization method can make it more easily to identify intellectual turning points by visual inspection. The superstring dataset in this study is updated to include citation data between 1985 and 2003.

Visualized networks were validated by the leading scientists in the field of superstring. We showed the merged map, without pruning, to John Schwarz (California Institute of Technology, Pasadena) and Edward Witten (Princeton University, Princeton). Schwarz is the coauthor of the article that triggered the first superstring revolution; Witten has written a number of highly cited articles on superstring and is also the top-ranked physicist in a list of the 1,000 physicists most cited between 1981 and 1997. The list was compiled by the Institute for Scientific Information, who were asked to explain the nature of intellectual contributions identified by pivot points and hubs in the networks.

The 19-year time interval was sliced into six 3-year segments, starting from 1985–1987 and ending with 2000–2002, plus a 1-year segment for 2003. Two sets of results were generated from two separate runs: one used relatively higher-threshold settings, which resulted in small networks (Table 1); the other used lower-threshold settings for larger networks (Table 2). Two versions of the larger network are shown: one without global pruning (Fig. 3) and the other with global pruning (Fig. 4). Links were color-coded by the earliest establishment rule. Darker colors indicate links from earlier time slices, whereas lighter colors indicate links from more recent slices. Networks in individual time slices are not shown due to page limitations.

Results

Table 1 shows the size of the cite space and details of individual networks and the merged network. The size of the cite space in a given time slice is the number of articles that have at least one citation within the given time slice; the size is generally increasing over time. The size for 2003 is smaller, because the 2003 data are still incomplete. The merged network contains 82 articles, and various pivot points are evident at a glance (Fig. 2).

Table 2 shows the threshold setting for a sequence of larger networks. The cocitation network in each time slice represents approximately the top 1% most cited articles. The merged network contains 647 unique articles, which collectively made 1,097 appearances in these time slices. In other words, 41% of articles appeared in more than one time slice. The locally pruned version of the merged network is shown in Fig. 3; the globally pruned version is shown in Fig. 4.

As shown in Fig. 3, color-coded links in effect partitioned the merged network into several major clusters of articles. Clusters of the same color represent cocitations made within the same time slice. More importantly, as we expected, within-cluster cocitation links are evidently more common than between-cluster links. A strongly clustered network also makes it easy to identify pivot nodes and between-cluster links. Six structurally strategic nodes are identified in Fig. 3, including the 1984 Green–Schwarz article, which triggered the first superstring revolution. However, the 1995 Polchinski article that triggered the second superstring revolution was not obvious in the dense visualization; Polchinski introduced the fundamental concept of D-branes in that article.

The 1984 Green–Schwarz article is a typical pivot node; it is the only contact point between two densely connected clusters in blue (1985–1987). It was this article that sparked the first superstring revolution, the famous 1984 Green–Schwarz anomaly cancellation paper. Friedan's 1986 article is a distinct pivot node connecting blue (1985–1987), pink (1988–1990), and green clusters (1991–1993). Witten's 1986 article is a pivot between a

Table 2. Time slicing and threshold settings for a set of larger networks, where f_c is the citation frequency threshold and f_{cc} is the cocitation frequency threshold

Time slices	f_c	f_{cc}	Cite space size	Top cited	Sample, %	cc (cosine ≥0.15)
1985–1987	2	1	604	39	6.46	229
1988–1990	4	3	2,740	114	4.16	283
1991–1993	15	7	12,214	200	1.64	1,263
1994–1996	20	10	16,147	229	1.42	895
1997–1999	25	10	19,716	223	1.13	956
2000–2002	30	15	22,449	180	0.80	486
2003	10	10	9,594	112	1.17	131
Total (unique)			83,464	1,097 (647)	Mean 2.4	Total 4,243

cc, cocitation.

blue cluster (1985–1987) and a yellow cluster (2000–2002). Fig. 3 also contains a couple of smaller clusters that are completely isolated from the main super cluster. Small clusters in red (2003) indicate the candidates for emerging clusters. We were able to find Polchinski's 1995 article in a smaller-sized merged network, but the article must be overwhelmed by the 4,000 strong links of the larger network. Nevertheless, the quality of the visualized network is promising: intellectually significant articles tend to have topologically unique features.

Articles by Maldacena, Witten, and Gubser–Klebanov–Polyako, located toward the top of the major network component, were all published in 1998. When we asked Witten to comment on an earlier version of the map, in which citation counts were not normalized by years since publication, he indicated that the Green–Scharwz article is more important to the field than the three top-cited ones, and that the earlier articles in the 1990s appeared to be underrepresented in the map. The apparent mismatch between citation frequencies of nodes and their importance judged by domain experts was partially corrected in the network shown in Fig. 2. Witten's comments raised an important question: is it possible that an intellectually significant article may not always be the most highly cited?

Fig. 4 shows the merged network pruned by Pathfinder; the pruned version contains fewer links than the version in Fig. 3. Much of the within-cluster links is reduced to links between cluster centers and other cluster members. Links between non-center members are essentially removed. The overall structure is simpler and easier to explore. In addition, the number of link colors attached to a node distinguishes a pivot node from a nonpivot node. If a node connects to other nodes through links in a single color, it is not regarded as a pivot node, because it does not imply intellectual transitions over time. In contrast, if a node joins several different-colored links, it is a good candidate for an intellectual turning point, because if paths connecting articles in different clusters must go through a pivot point, the pivot point is likely to have a unique position in the literature.

The Green–Schwarz article is located toward the center of the visualization; it joins links from four different time slices. The 1995 article by Candelas *et al.* is similar in terms of the link colors. According to Institute for Scientific Information's Sci-

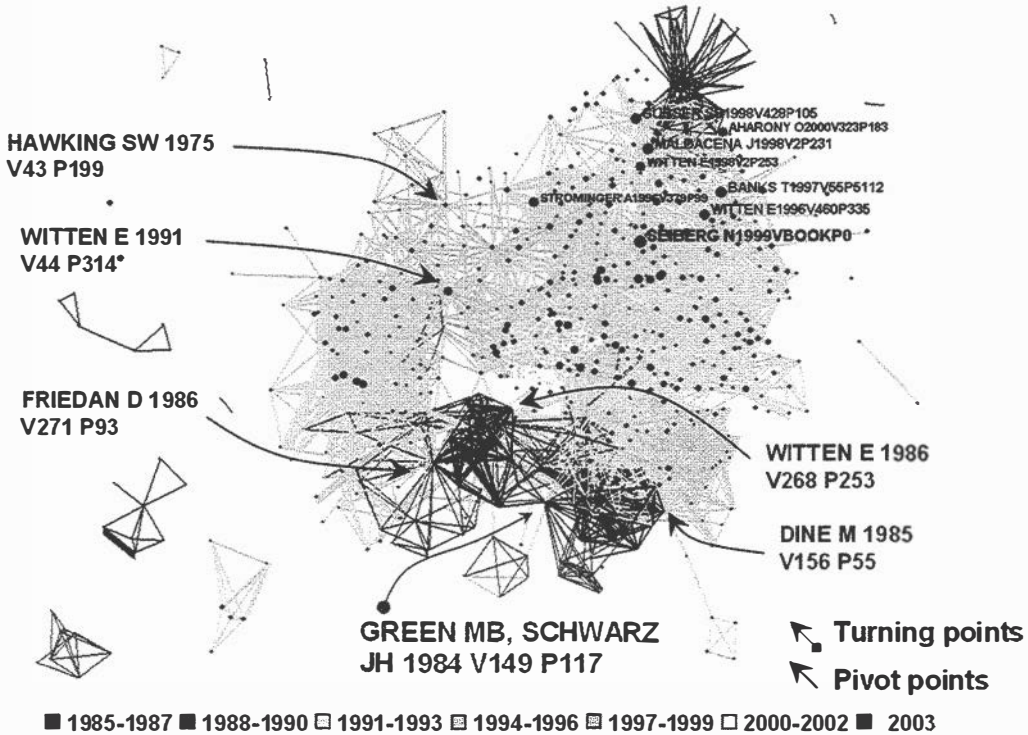


Fig. 3. A 624-node merged network without global pruning. See a color version at: www.pages.drexel.edu/~cc345/citespace/Figure3.png.

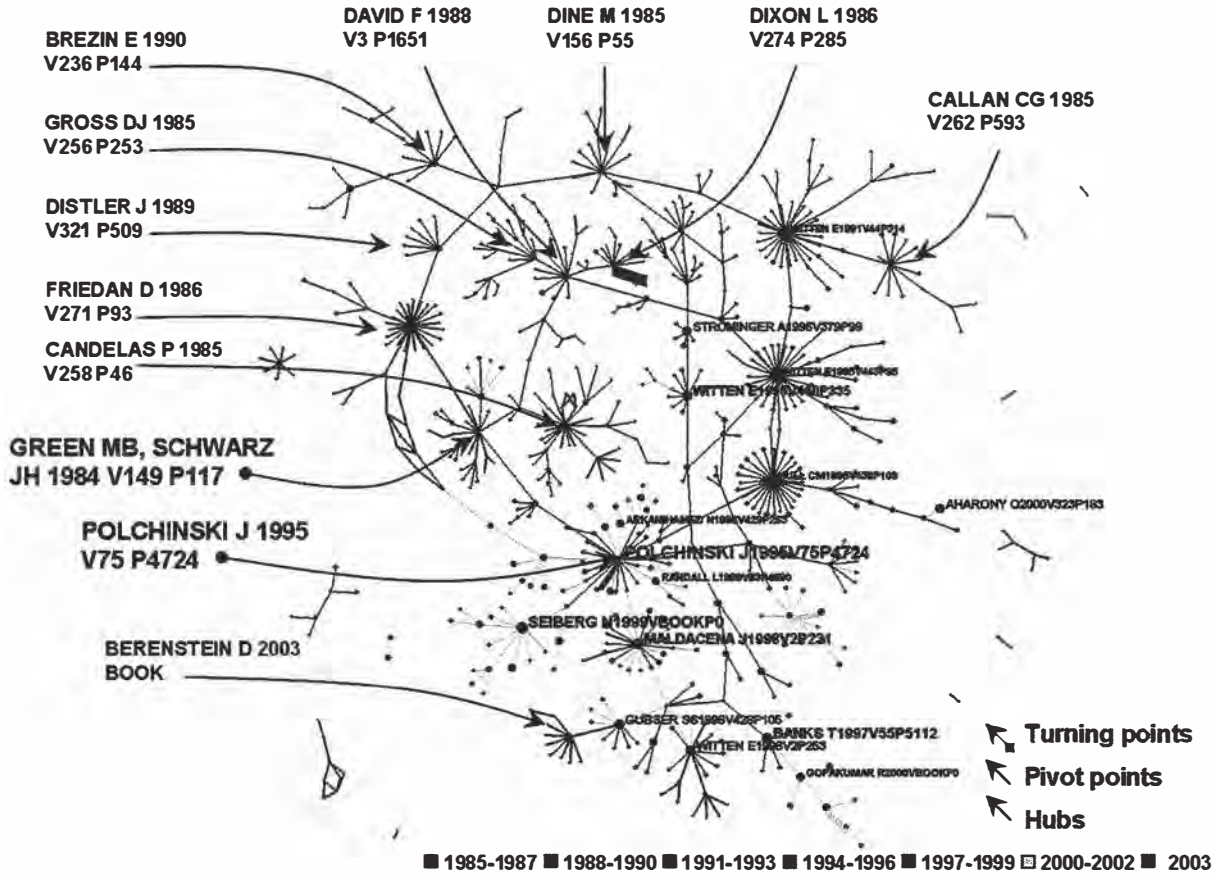


Fig. 4. A 624-node merged network with global pruning by using Pathfinder ($q = N - 1$, $r = \infty$). See a color version at www.pages.drexel.edu/~cc345/citespace/Figure4.png.

ence Citation Report (1981–1998), Candelas *et al.* 's article has a total of 1,538 citations during that period, and its average annual citation is 110. The 1995 Polchinski article can be easily found at the lower center of the map; according to Schwarz, it explained the concept of D-branes, a crucial ingredient in almost all modern string theory research. It appears to be more of a hub than a pivot node, connected by links of only two colors, brown (1997–1999) and yellow (2000–2002). To the left of Polchinski's article is a cluster centered by the 1998 Maldacena article. Schwarz noted that in this article, Maldacena made a major new discovery that in certain circumstances relates string theories to quantum field theories.

The comments from domain experts have confirmed that both versions of the merged network indeed highlight significant articles, and these articles tend to have unique topological properties that distinguish them from other articles. The globally pruned version is easier to explore than the local-pruning-only version.

Discussion

The results are particularly encouraging because the presence of pivot nodes enables us to narrow down the visual search quickly to a small number of good candidate nodes for intellectual turning points. An easy identification of such turning points is an important and necessary step toward effective detection of paradigmatic changes in a knowledge domain. The small network is particularly clear, containing both turning points. The larger network without pruning is cluttered, although it is still possible to identify several pivot points.

The interpretation and validation of the visualizations have greatly benefited from help from leading scientists in the knowledge domain. The work has also shown that using a variable threshold could be a potentially good practice for citation analysis in general.

In comparison to our earlier visualization of the superstring cocitation network (8), this method tends to produce more distinct visual features for key articles. More importantly, such visual features appear to be independent to the amount of citations of a node. In other words, a lower citation rate is not necessarily preventing a node from having salient visual features, suggesting that cocitations must have played a greater role. Pivot nodes can be identified even if they have relatively fewer citations. This could be a particularly useful feature for the detection of significant articles that could be easily overlooked by falling below a single high-citation threshold.

The 1984 Green–Schwarz article for the first revolution is a typical pivot node, whereas the 1995 Polchinski article for the second revolution is more of a hub than a pivot node. This finding suggests that before we have further evidence, it would be sensible to examine both types of visualizations, pruned and unpruned, in a study of intellectual turning points.

In comparison to other methods for detecting changes of networks over time, our approach simplifies cognitively demanding tasks of comparing a sequence of network snapshots. The progressive visualization method allows us to focus on much simpler tasks of locating pivot nodes and cluster centers. The color-coded links enables the user to trace temporal patterns through the network visualization.

The progressive visualization method introduced here has practical implications. It provides scientists with a roadmap of their own field. Witten commented, "It was fun to look at it." A longstanding challenge is to be able to visualize cocitation networks of a domain as quickly as new bibliographic data become available so that one can monitor the changes of a domain more closely on a monthly or even weekly basis. The approach provides a practical starting point. Users have the flexibility to slice a time interval into smaller as well as larger segments.

Using overlapped time slices could be a valuable alternative to explore in future studies. Currently, adjacent time slices are mutually exclusive to highlight the magnitude of a potentially important change, whereas overlapping slice segments may blur such changes and make them less obvious to detect.

An unsolved issue is concerned with the detection of abrupt changes in citations within a short period. We normalized the citations of an article by its publication age. Additional metrics of pivot nodes should augment the power of visual inspection even further. Knowledge discovery and data-mining techniques, such as Kleinberg's burst-detection technique (35), are expected to play a substantial role in identifying a paradigm shift.

Finally, the role of domain experts in KDViz needs to be further investigated. Experts in the fields are the best sources to seek validations and interpretations. On the other hand, one should also use domain visualizations with caution; and it should be made clear that algorithmically generated domain visualizations, however crafted, merely portray the complexity of an underlying domain from a limited perspective. If KDViz can stimulate scientists to look at their own field from a different perspective and pose new questions about the evolution of their domain, KDViz will ultimately become a practical tool to study science itself.

Conclusion

The progressive KDViz method simplifies the tasks of tracking significant changes of a knowledge domain's cocitation network over time. Cognitively demanding tasks of comparing complex networks back and forth are simplified to tasks of locating pivot points and cluster centers in visualized networks.

The divide-and-conquer strategy maximizes the strengths of algorithms and reduces the influence of their weaknesses. The cosine cocitation coefficients are effective enough to pick up the most intellectually significant articles, whereas the Pathfinder-enhanced version improved the quality even further.

CITESPACE provides an experiment platform to investigate new ideas and compare existing approaches. We plan to make a further refined version of CITESPACE available in the near future to researchers, practitioners, and educators in various disciplines and obtain their first-hand experience in capturing the changes of their own domains.

Further studies and in-depth case studies of progressive KDViz should be encouraged. For example, can this method detect the merge of two domains or the split of a single domain into a few new ones? Can this method detect scientific revolutions in other disciplines? Will it work with alternative representations of a knowledge domain, such as the preprint archives used by physicists and other sources? KDViz is a challenging route, but it is also potentially rewarding for scientists in so many different knowledge domains to have easy access to the big picture of their own fields.

We give special thanks to John Schwarz (California Institute of Technology) and Edward Witten (Princeton University) for help in interpreting the visualizations. The 2002 Institute for Scientific Information/American Society for Information Science and Technology Citation Analysis Research award is acknowledged.

1. Chen, C. (2003) *Mapping Scientific Frontiers: The Quest for Knowledge Visualization* (Springer, London).
2. Barabási, A.-L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A. & Vicsek, T. (2002) *Phys. A* **311,** 590–614.
3. Newman, M. E. J. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 404–409.
4. Garfield, E. & Small, H. (1989) in *Innovation: At the Crossroads Between Science and Technology*, eds. Kranzberg, M., Elkana, Y. & Tadmor, Z. (Neaman, Haifa), pp. 51–65.
5. Small, H. & Greenlee, E. (1989) *Commun. Res.* **16,** 642–666.
6. Kuhn, T. S. (1962) *The Structure of Scientific Revolutions* (Univ. of Chicago Press, Chicago).
7. Small, H. G. (1977) *Soc. Stud. Sci.* **7,** 139–166.
8. Chen, C. & Kuljis, J. (2003) *J Am. Soc. Inf. Sci. Technol.* **54,** 435–446.
9. Price, D. D. (1965) *Science* **149,** 510–515.
10. Albert, R. & Barabasi, A. (2002) *Rev. Mod. Phys.* **74,** 47–97.
11. Dorogovtsev, S. N., Mendes, J. F. F. & Samukhin, A. N. (2000) *Phys. Rev. Lett.* **85,** 4633–4636.
12. Newman, M. (2001) *Phys. Rev. E* **64.**
13. Havre, S., Hetzler, E., Whitney, P. & Nowell, L. (2002) *IEEE T. Vis. Comput. Graphics* **8,** 9–20.
14. Carroll, J. D. & Chang, J.-J. (1970) *Psychometrika* **35,** 283–319.
15. Gower, J. C. (1975) *Psychometrika* **40,** 33–51.
16. Bookstein, F. L. (1989) *IEEE T. Pattern Anal.* **11,** 567–585.
17. Brandes, U. & Corman, S. R. (2003) *Inf. Visual.* **2,** 40–50.
18. Erten, C., Harding, P. J., Kobourov, S. G., Wampler, K. & Yee, G. (2003) *Technical Report TR0304* (Univ. of Arizona, Tucson, AZ).
19. Misue, K., Eades, P., Lai, W. & Sugiyama, K. (1995) *J. Visual Lang. Comput.* **6,** 183–210.
20. North, S. C. (1995) in *Proceedings of Graph Drawing* (GD'95), ed. Brandenburg, F. J. (Springer, New York), pp. 409–418.
21. Batagelj, V. & Mrvar, A. (1998) *Connections* **21,** 47–57.
22. Ware, C., Purchase, H., Colpoys, L. & McGill, M. (2003) *Inf. Visual.* **1,** 103–110.
23. Schvaneveldt, R. W. (1990) *Pathfinder Associative Networks* (Ablex, Norwood, NJ).
24. Chen, C. (1999) *Inform. Process. Manag.* **35,** 401–420.
25. Chen, C. & Paul, R. J. (2001) *Computer* **34,** 65–71.
26. Chen, C., Cribbin, T., Macredie, R. & Morar, S. (2002) *J Am. Soc. Inf. Sci. Technol.* **53,** 678–689.
27. Chen, C., Kuljis, J. & Paul, R. J. (2001) *IEEE T. Syst. Man. Cy. C* **31,** 518–529.
28. White, H. (2003) *J Am. Soc. Inf. Sci. Technol.* **54,** 423–434.
29. van Raan, A. (2000) *Scientometrics* **47,** 347–362.
30. Ahlgren, P., Jarneving, B. & Rousseau, R. (2003) *J Am. Soc. Inf. Sci. Technol.* **54,** 550–560.
31. White, H. D. (2003) *J Am. Soc. Inf. Sci. Technol.* **54,** 1250–1259.
32. Chen, C. & Morris, S. (2003) in *IEEE Symposium on Information Visualization* (InfoVis'03) (IEEE Computer Society Press, Seattle), pp. 67–74.
33. Kamada, T. & Kawai, S. (1989) *Inform. Process. Lett.* **31,** 7–15.
34. Schwarz, J. H. (1996) arXiv:hep-th/9607067 (http://arxiv.org/PS_cache/hep-th/pdf/9607/9607067.pdf).
35. Kleinberg, J. (2002) in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York), pp. 91–101.

Arthur M. Sackler
COLLOQUIA
OF THE NATIONAL ACADEMY OF SCIENCES

Mapping Knowledge Domains

May 9–11, 2003

Arnold and Mabel Beckman Center, Irvine, CA
Organized by Richard M. Shiffrin and Katy Börner

Program

Friday, May 9

Keynote Address

Eugene Garfield, ISI
A Historiograph of Mapping Knowledge Domains

Session I: Data Bases, Data Format, and Access

Monika Henzinger, Google, and Steve Lawrence, NEC Research Institute
Extracting Knowledge from the World Wide Web

Paul Ginsparg, Cornell University
Mapping Subsets of Scholarly Information

Saturday, May 10

Session II: Data Analysis Algorithms

Thomas K. Landauer, University of Colorado at Boulder and Knowledge Analysis Technologies
From Paragraph to Graph

M. E. J. Newman, University of Michigan
The Structure of Scientific Collaboration Networks

Elena Erosheva, University of Washington
Using Mixed Membership Models for Mapping Knowledge Domains

Thomas L. Griffiths, Stanford University, and Mark Steyvers, University of California, Irvine
Topic Dynamics in Knowledge Domains

Jon Kleinberg, Cornell University
Enhancing Web Sites with Usage Data

Kate McCain, Drexel University
Combining Bibliometric and Knowledge Elicitation Techniques to Map a Knowledge Domain

Sunday, May 11

Session III: Visualization and Interaction Design

Colin Ware, University of New Hampshire
Information Seeking and the Objects of Visual Attention

Alan M. MacEachren, Pennsylvania State University
Geovisualization for Constructing and Sharing Concepts

Chaomei Chen, Drexel University
Paradigms, Debates, and Puzzles in Science: A Visual Exploration

Katy Börner, Indiana University
The Simultaneous Evolution of Article and Author Networks in PNAS

Session IV: Promising Applications

Susan Dumais, Microsoft Research
Visualizing Search Results

Beth Hetzler, Pacific Northwest National Laboratory
Analysis Experiences Using Information Visualization

Kevin W. Boyack, Sandia National Laboratories
An Indicator-Based Characterization of the *Proceedings of the National Academy of Sciences*

Francis Narin, CHI Research, Inc.
From Science Papers to Technology Patents and on to Company Financial Performance

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce M. Alberts is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Wm. A. Wulf is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Bruce M. Alberts and Dr. Wm. A. Wulf are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org