Discovering Communities through Information Structure and Dynamics: A Review of Recent Research

Nick Montfort

Technical Report MS–CIS–04–18
Department of Computer and Information Science
University of Pennsylvania

9 August 2004

Contents

Introduction

Social organizations have important structures that are not immediately evident. Rather than being monolithic or completely arranged in neat hierarchies, organizations of all sorts — neighborhoods, universities, businesses, international academic associations, political movements, and others — contain numerous different communities that are not described on any formal org chart or in any official directory. For instance, a university department, in addition to being organized into official labs and research groups, also may be organized into shorter-term teams working on particular research papers; teams working together to develop and grade homework assignments and exams; communities of practice (Matlab programmers, Perl programmers, etc.); communities of interest (squash players, aficionados of vampire films, etc.); and other sorts of groups, none of which will be described in formal documents about the department. Yet the true structure of an organization is an important factor in determining how information diffuses within it, how consensus can be built, and which people are essential to accomplishing which of the organization's objectives.

Since some of the discourse that forms, maintains, and indicates communities travels through networked computers, an analysis of computer-mediated communications can shed light on the nature of these organizational structures. Even if the content of such communications is not examined for reasons of privacy or practicality, there is much to be learned from the information flow within an organization and from the hypertext structure of related pages on the Web. This report considers three techniques for community discovery on the Web [10, 9, 2] and the ways that two groups of researchers have more recently partitioned email graphs into communities.[26, 8]

1.1 The Notion of "Community"

When seeking to discover communities, it is important to define *community,* or at least to characterize this concept so that performance on the task of community discovery can be measured. Community is a more specific concept than *organization,* defined as "any stable pattern of transactions between individuals or aggregations of individuals."[21, p140]

Traditional sociological definitions of community have emphasized physical proximity and the sharing of tangible resources — seemingly in contradiction to the idea of an online community, although common goals were also often mentioned in early definitions.[14] Before the Web had surged in popularity, Howard Rheinghold championed the term *virtual community* to describe "social aggregations that emerge from the Net when enough people carry on public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace."[22, p5] The concept of a community that was connected by networked computers existed long before the 1990s, however. In 1968, J. C. R. Licklider and Robert Taylor wrote:

> What will on-line interactive communities be like? In most fields they will consist of geographically separated members, sometimes grouped in small clusters and sometimes working individually. They will be communities not of common location, but of *common interest.* In each field, the overall community of interest will be large enough to support a comprehensive system of field-oriented programs and data.[20]

Licklider was prescient in seeing that people who do not dwell in the same geographical area can be members of the same community. Although telephones, the highway system, and air travel had provided ways for people to maintain long-distance social ties before Licklider described on-line communities, communities were typically considered to be local groups through the 1950s and 1960s.

Since the late 1960s, sociologist Barry Wellman has been studying communities as non-local social networks. He has provided an updated definition of community that is sensitive to current social realities, including our social uses of the networked computer: "Although community was once synonymous with densely knit, bounded neighborhood groups, it is now seen as a less bounded social network of relationships that provide sociability support, information, and a sense of belonging."[27, p2031] Further defining may be necessary before we know whether the vertex corresponding to a particular person should be connected to a particular community graph, but by developing a consistent standard for a community relationship, based upon this fundamental definition, a principled way to model communities can be developed. This definition seems suitable for locating communities within a variety of organizations, including companies, universities, cities, and the Internet.

When researchers' definitions of terms such as *community* and *friend* differ substantially from the ones above or from intuitive ones, the differences are discussed in section 2 as the corresponding research is summarized. The implications of these differences are explored in section 3.

1.2 Social Communication and Networked Computing

The networked computing environment at the beginning of the 21st century is dominated by the Internet, which carries the bulk of our computer-mediated communications. The Internet's predecessor, ARPANET, was largely envisioned by Licklider, who imagined a system to facilitate resource sharing and collaborative work.[19] Given the Internet's origins, it is hardly surprising that this network has helped to bring about new and enhanced forms of social communication. The Internet has turned out to be, as Wellman wrote:

> ... an excellent medium for supporting far-flung, intermittent, networked communities. E-mail transcends physical propinquity and mutual availability; e-mail

> lists enable broadcasts to multiple community members; attachments and Web sites allow documents, pictures, and video to be passed along; buddy lists and other awareness tools show who might be available for communication at any one time; and instant messaging means that simultaneous communication can happen online as well as face-to-face and by telephone.[27, p2031]

In the past few decades, observers of computer-mediated communication made distinctions between online or virtual communities and "real life" — or "non-simulated reality," as it was called in Richard Powers' novel *Plowing the Dark.* Opponents of computer-mediated communication, recapitulating concerns about urbanization and the development of pre-computer technologies, have worried that "real life" community will be forgotten or disrupted. Proponents of the networked computer often seemed to admit that online communities were separate from real ones, but argued that they are just as good in many ways, and that participation in them does not disturb "real" community life.

Such arguments, on both sides, miss the point that many communities cannot be simply classified as online or offline, since members communicate with each other both with and without computer mediation. This sort of discussion also gives second-class status to social ties established online, presupposing that the loss of a social tie in the "real world" is always a horrible outcome, even if a person lost that tie while establishing one or more social ties online, perhaps ties that are even more meaningful and important. The Internet, according to Wellman, "is not destroying community but is resonating with and extending the types of networked community that have already become prevalent in the developed Western World."[27, p2032] The ability to communicate online can help families keep in touch across great distances, just as it can help a business that contracts with companies in other counties, a team of scientific researchers at different universities and research labs, and a team of editors living on different continents.

While dispersed, non-local communities are enabled by the networked computer and make for interesting and important objects of study, Wellman's research has shown that "online as well as offline contact is highest with those living nearby;" he cites other findings indicating that a significant number of computer-mediated communications are undertaken to arrange non-computer-mediated meetings.[27, p2033] This suggests that an analysis of computer-mediated communications can shed light on the structure and development of geographically localized communities as well as distributed ones.

1.3 Social Network Analysis

Social network analysis describes a set of methods that have been developed since the middle of the last century by researchers who have drawn on sociometrics, graph theory, and other theories and techniques to try to better understand human relationships and

communications.[24] Management and organizational research has increasingly found techniques developed in this field to have practical value in understanding human networks and improving the functioning of businesses. One review of recent organizational research found that this field's increase in social network research, over the past three decades, has been exponential.[3]

Social network analysts use questionnaires and surveys, documentary sources, and observation to gather quantitative data about *attributes* (statistics of individuals, such as income), *relations* (which determine the structure of the graph), and *ideations* (for instance, people's motives). The structure of the network is of primary concern. The *centrality* of the people in the network is important in this analysis; different measures of centrality are used. One of these is simply *high degree,* a local measure of centrality. Another measure, this one global, is *closeness,* which is proportional to the number of vertices that can be reached and is inversely proportional to the length of the shortest paths to those vertices. In other words, vertices that are topologically near many other vertices have high closeness. In a connected graph, the closeness of a vertex is just one over the sum of its distances to each other vertex.

Malcolm Gladwell popularized some social network analysis concepts in a book that discusses the spread of ideas.[12]. Gladwell described three sorts of people critical to an idea's diffusion: the *maven,* who is knowledgeable and likes to help others with information; the *salesman,* who is particularly persuasive about ideas; and the *connector,* who significantly shortens the paths of acquaintanceship between people, sometimes by virtue of having high degree and sometimes because this person bridges two or more subgraphs that would otherwise be disconnected, or connected only by very long paths. The special qualities of mavens and salesmen are best described in attributional and ideational ways. Connectors, on the other hand, can be identified with reference to graph structure, as vertices with high *betweenness.* The betweenness of vertex v is related to the number of all-pairs shortest paths passing through v. More precisely, all shortest paths that include v, between each pair of vertices s and t, contribute to the betweenness of a vertex. The contribution of each pair is the number of shortest paths that include v divided by the total number of shortest paths between the pair. If all the shortest paths for s and t pass through v, the contribution is 1; if one of three shortest paths pass through it, the contribution is $1/3$. Betweenness, a global measure, is a third measure of centrality.

The methods of social network analysis include *cluster analysis,* techniques for identifying highly interconnected subgraphs. These are of particular interest in the community identification task.

Recent Research on Community Discovery

There has been significant recent research on determining communities mainly or purely through the structure of the graph of online communications. While some research has considered communities in the context of USENET newsgroups[16, 25, 23] or chat systems[17], interesting work has been also done using data from two of the most popular means of computer-mediated communication, the World Wide Web and email. These two systems differ in important ways. Pages on the Web are public and somewhat loosely associated with individual people, while email messages are private and tightly bound to individuals. A link on the Web usually represents an intentional reference, but the "recipient" may never even know about the link, as is seldom the case with an email that is sent. Because of these differences, and for other reasons, approaches to community identification in these two systems have varied.

2.1 Using Web Structure to Find Communities

Parts of the Web have been modeled as graphs in several different ways, but all of the work described below considers Web pages to be vertices and hyperlinks to be edges, usually directed ones. Although much work using this sort of model focuses on the identification of pages relevant to a particular topic, the research surveyed in this section explicitly concerns itself with the identification of communities or social relationships. The analysis of citation structures (bibliometrics) is sometimes referred to and used as a point of comparison in this research, but it can be understood without reference to particular bibliometric techniques.

2.1.1 Determining Communities with HITS

Three researchers working at IBM Almaden Research Center and at Berkeley developed a technique for community identification [10] based on the Hypertext-Induced Topic Search (HITS) algorithm developed by one of them, Jon Kleinberg.[18] A set of pages that have been returned from a search engine is expanded and numerous different communities are identified within the enlarged graph, with the most important pages in these communities indicated. David Gibson, Jon Kleinberg, and Prabhakar Raghavan do not define *community,* but they note that the term does not imply centralization or planning. They suggest a connection to the social world by noting that an analysis of Web structure "gives us a global understanding of the ways in which independent users build connections to one another in hypermedia that arises in a distributed fashion."[10, p225]

HITS was developed for broad-topic searches. It uses the mutually-defined concepts of *hubs* and *authorities*: hubs are pages that point to many authorities, authorities are pages pointed

to by many hubs. Vertices with high out-degree and high in-degree are not the same things as hubs and authorities, although these turn out to be the first-iteration approximations. At a high level, the algorithm works by starting with a *root set* of pages (which were returned as a search result) and then expanding the set, adding all pages that link to or are linked to by ones in the root set. Each vertex v is assigned an authority weight x_v of 1 and a hub weight y_v of 1. Consider Γ_v^- to be the neighborhood of vertices that point to v and Γ_v^+ to be the neighborhood of vertices that are pointed to by v. At each iteration, the weights are updated as follows:

$$x'_v \leftarrow \sum_{u \in \Gamma_v^-} y_u \qquad\qquad y'_v \leftarrow \sum_{u \in \Gamma_v^+} x_u$$

The hub weights are then normalized so that the sum of their squares is one, and similarly with the authority weights.

The algorithm above differs in only one way from the one used by Kleinberg and his colleagues. They did not expand the root set by fetching every page that pointed to pages in the set. Some popular pages were pointed to by hundreds of thousands of others, so if there were more than 50 pages in Γ_v^-, 50 of those were selected at random and added. The algorithm was observed to converge quickly; in experiments, 200 iterations would reliably result in convergence.

Let A be the directed adjacency matrix corresponding to the expanded set of pages, with $A_{ij} = 1$ if there is a directed edge from i to j, and 0 otherwise. Now the update operations can be described simply as $x \leftarrow A^T y$ and $y \leftarrow Ax$. For a large number of iterations k, the x and y vectors converge in the direction of $(A^T A)^k z^{(1)}$ and $(AA^T)^k z^{(1)}$, where $z^{(1)}$ is the initial vector of ones used to initialize both hub and authority weights. This reformulation supplies a proof of the convergence of the algorithm, since the matrix problem converges. Kleinberg cites a standard linear algebra result [13], that for any symmetric $n \times n$ matrix M and any n-vector v not orthogonal to $\omega_1(M)$, the direction of $M^k v$ converges to that of $\omega_1(M)$ as k grows without bound. Solving for the leading eigenvector (or an approximation of it) can in general be done in this way; this is called the *power method.* Also, $z^{(1)}$, a vector of all ones, cannot be orthogonal to $\omega_1(AA^T)$, because of another standard result, that the principal eigenvector of such a matrix has only nonnegative components. So the iterative process, equivalent to finding $(AA^T)^k z^{(1)}$ for large k, converges for the hub weights. A similar argument applies for $\omega_1(A^T A)$ and the authority weights. In addition to providing a convergence proof, this reformulation also offers a helpful correspondence that can be used to find the hubs and authorities of different "modes" within a subgraph. As k goes to infinity, $(A^T A)^k z^{(1)}$ approaches ω_1, the eigenvector that corresponds to the first (greatest) eigenvalue λ_1 of $(A^T A)$, and similarly for the update of the hubs weights and (AA^T). Following Kleinberg, this assumes that the multiplicity of λ_1 is always 1, so ω_1 can be called the *principal eigenvector* and the others *non-principal.* Kleinberg writes that "When the assumption does not hold, the analysis becomes less clean, but is not affected in any substantial way."[18, p613] The power method only works to find the leading eigenvector when there are n distinct eigenvectors and eigenvectors with different norms; if these conditions do not hold, convergence is not guaranteed. The analysis of the iterative algorithm is, therefore,

certainly affected when λ_1 has multiplicity greater than 1. However, other techniques can be used to determine the eigenvalues and eigenvectors even when the iterative algorithm is not guaranteed to converge, and the results can be interpreted in a similar way, with the eigenspace corresponding to λ_1 playing the same role as ω_1 does in the simplified case.

The principal eigenvector and the non-principal eigenvectors are interpreted as corresponding to principal and non-principal communities that have been found in the expanded set of pages.[18, 10] Given a particular authority eigenvector ω_k, the components of that eigenvector that have the greatest absolute values are interpreted as corresponding to the most authoritative nodes in the k^{th} community, and similarly for a hub eigenvector.

The main communities tended to recur when the root set was varied in ways that preserved its relevance to a single topic. The principal community did not always remain the same with different root sets, but the top five authorities overlapped significantly whether one began with a query on *astrophysics, astrophysique,* or *astrophysik,* for instance. HITS functions to generalize topics that are not sufficiently broad, finding hubs and authorities relevant to a less specific topic, but one that includes the topic originally chosen. This generalization was seen to work differently for seemingly "parallel" topics, such as *English literature* (top authorities were focused on the topic) and *German literature* (top authorities were more general, and relevant to European literature.) It was suggested that this was due to differences in the development of Web communities on such parallel topics, which implies that HITS can be used to gauge the development of such communities. Additionally, it was suggested that a hierarchy of topics, such as the one seen in Yahoo!, could be developed using the generalization ability of HITS.[10] Given a large set of highly specific topics and a common root set, the topics that generalized to the same community would all be considered hierarchically underneath the generalized topics. For instance, if running HITS on a dozen researchers' names all generalized to the same community, associated with machine learning, those names would be placed beneath "Machine Learning" in a hierarchical index.

Since "what determines the 'generality' of a topic in this setting is its *representation* on the WWW"[10, p231], some topics can be specialized rather than generalized by HITS, due to a strong presence of pages about a subtopic. For instance, the top two authorities in a root set expanded from a search on *linguistics* were both from a more specific field, *computational linguistics.* Although disappointing when considered as a search result, this seems to indicate that in 1998, the community of computational linguists as expressed on the Web was the dominant community dealing with "linguistics" overall. Highly-commercialized pages appear frequently for some topics, which could be due to the engineering of link structures and page text — this would influence the way the root set is selected. It could be caused by pages propagating weights within a domain, and could suggest that such weight propagation be limited or eliminated. The presence of AltaVista and Yahoo! as highly-ranked authorities in many fields suggests their infiltration into many communities. The influence of short-term changes (conferences, events reported in the news) on the top authorities for a particular topic suggested that the "core" of a topic could be determined by superimposing HITS authorities determined over a long time period and then selecting the ones that remain constant.[10]

The different communities indicated by different eigenvectors may correspond to different senses of a search term or to different, seldom-interlinked takes on the same topic. There is also some suggestion, based on work on heuristics for spectral partitioning of undirected graphs[5], that the sign of an eigenvector's component (corresponding to a particular authority or hub weight) may be significant, since vertices whose components have opposite signs have been observed to be well-separated in undirected graphs. Examples were given of different communities found with root sets based on searches for *jaguar, randomized algorithms,* and *abortion.*[18] The top three "jaguar" authorities corresponded to the Atari game console (in the principal eigenvector), the Jacksonville football team (in the second non-principal eigenvector), and the automobile (in the third). With "randomized algorithms," the first non-principal community had positive-valued authorities that were the home pages of computer scientists and negative-valued authorities that were software packages. The second non-principal eigenvector for the "abortion" set provided a clear separation of a different sort, with the top six positive-valued authorities all pro-choice pages and the top six negative-valued ones all pro-life.

2.1.2 Community Identification as a Maximum Flow Problem

Researchers at NEC developed a technique for approximately identifying a community by creating an approximate minimum cut.[9] Gary William Flake, Steve Lawrence, and C. Lee Giles defined a *community* as a set of Web pages that has at least as many edges (in either direction) to pages within the set as it does to pages outside the set. In practice, this determination was made on a subgraph of the Web found by crawling three links out from a particular page. The problem was initially formulated as a balanced minimum cut problem: remove a set of edges such that the number removed is minimal, while ensuring that the two disconnected subgraphs remaining have at least m vertices each.

Computing this balanced minimum cut exactly is NP-complete. Flake *et al.* reformulated it as a maximum flow problem. Source s is the initial page of the crawl, always inside the community, which has c vertices in it, and t is a "virtual sink," added to the pages most distant from s in the graph that results from the crawl. A proof is given that with this "virtual sink," and with an additional condition, if the edges to the virtual sink are given capacity 1 and all other edges given capacity k, the same cut set will be found as would happen if a real vertex, located outside the community, were used as the sink. Call the number of vertices that are not in the community to be l, and the capacity of the original cut set c. Then, the additional condition can be stated as $1 < k < \frac{l}{c}$. The proof proceeds constructively, in four steps, beginning with the original situation and transforming into the one with a virtual sink, showing that the result is unchanged at each step: (1) Multiplying all edge capacities in the original graph by k changes nothing about the solution; (2) Connecting all non-community vertices to the virtual sink with edges of capacity 1 leaves the cut where it as, as long as there are more than ck non-community vertices, and this is part of the additional condition $1 < k < \frac{l}{c}$; (3) Connecting all community vertices to the virtual sink with a capacity 1 edge also doesn't change the solution, since $1 < k$ implies that cutting

that new edge will be more efficient than cutting one of capacity k, and (4) The trivial cut in which all edges to the virtual sink are removed is shown to be more expensive than the original cut, since $k < \frac{l}{c}$. However, as discussed in section 3.1.2, the final technique that was used was not the one for which this property was proved.

Three community subgraphs were identified and member pages were given a score within each that was the sum of their in-degree and out-degree, considered within the community subgraph, not within the entire Web. Beginning with four seeds, a community based around support vector machines was identified; the highest scoring page was the home page of Vladimir Vapnik, who initially developed SVMs. A community associated with the Internet Archive was identified by beginning with eleven seeds. Finally, a "Ronald Rivest community" was identified beginning with a single URL, Rivest's home page, and allowing links internal to `mit.edu` to be traversed on the first crawl. (Other crawls did not follows links within the same domain.) A relevant community was identified, with two of three top scorers being Rivest's co-authors on the first edition of *Introduction to Algorithms.*

2.1.3 Finding Friends through Text and Link Similarity

Research done at Xerox PARC [2] explored how well links between student home pages could be predicted. Lada Adamic and Eytan Adar used three types of data: links (both to and from home pages), named entities in the text of home pages, and mailing list membership. Co-occurrences of items of each type were used to determine how similar pages were. The similarity was simply a weighted sum of items that were shared between pages. The similarity weight provided by each item was adjusted so that items shared by larger numbers of people had a lower weight; the weight assigned to an item was 1 over the log of the number of people sharing that item. Specifically, $similarity(A, B) = \sum_{i \in A, i \in B} \frac{1}{log(count(i))}$. This means that common named entities (e.g., "Electrical Engineering") contribute less than do unusual ones, (e.g., "NTUA," indicating the National Technical University of Athens), and similarly for links and mailing list memberships.

The goal was to predict whether people were "friends." Somewhat anticipating Friendster, these researchers used the term *friend* to simply mean "any user who links to or is linked to by another." Researchers evaluated which of these types of items were most predicative of "friendship," considering in-links and out-links as separate classes so that four classes of items were evaluated against each other. For both Stanford and MIT, in-links were the most predictive, then mailing list memberships, then out-links. Co-occurrences of named entities were the worst at predicting a link between sites. The specific links and terms that were the best predictors revealed some differences between Stanford and MIT: living groups and religious groups topped the list at MIT, where students often stay in one residence for four years, while the most predictive links, phrases, and mailing lists at Stanford were more heterogeneous.

2.2 Using Email Flows to Find Communities

Since email exchanges intuitively look more like conversations or correspondences than like journal articles that cite one another, it is unsurprising that bibliometric techniques are mentioned less often in considerations of email flows; interpersonal methods from social network analysis tend to be more directly applied to email graphs. The studies considered here look only at the information recorded in server logs, including "To:," "From:," and "Date:" lines but not including the message text itself.

2.2.1 Partioning the Email Graph with Betweenness

Three researchers from Hewlett Packard Labs developed and evaluated a technique for discovering community structure using an organization's email graph. Defining communities of practice as "the informal networks of collaboration that naturally grow and coalesce within organizations," [26, p82] Joshua Tyler, Dennis Wilkinson, and Bernardo Huberman used the centrality measure of *betweenness* to repeatedly partition an email graph. The different partitions were then aggregated. Subgraphs that appeared in many of the different partitions were identified as communities of practice.

The method of partitioning an undirected graph using betweenness had been introduced in a more general context by Girvan and Newman in 2002, who evaluated it on computer-generated networks and on real-world biological and social networks where the community structure was known.[11] The method was introduced as an alternative to a hierarchical technique that was traditionally used in cluster analysis but which produced poor results in many cases. The centrality measure of betweenness, introduced in section 1, is used here with reference to edges rather than vertices. Edges are of high betweenness if they are on many shortest paths. The basic algorithm is:

```
calculate betweenness for all edges in the graph
while edges remain in the graph do
   remove the edge with the highest betweenness
   recalculate betweenness for all edges in the affected component
end while
```

The different graphs that result at each iteration represent community graphs with different thresholds for the maximum permissible betweenness. A different **while** condition could be used so that the algorithm halts when a stopping condition is met; this is what was done by Tyler *et al.*

Let $n = |V|$ and $m = |E|$ for a graph $G = (V, E)$. Calculating the betweenness of each edge has traditionally been done with algorithms that take $\Theta(n^3)$ time and use $\Theta(n^2)$ space. The computation is dominated not by finding the shortest paths between all pairs, but by the final

summation; for each vertex v and each pair of vertices, it is necessary to compute what ratio of shortest paths v lies upon. In unweighted graphs, breadth-first search can be augmented to find the length and number of all paths from a given vertex in $O(m)$ time. However, the betweenness of a given vertex v must then be computed as the sum of the number of shortest paths between s and t that include v divided by the total number of shortest paths between s and t. Computing this sum involves going through every combination of s, t and v for $s \neq t \neq v$, so doing this directly takes $\Theta(n^3)$ time.

Tyler *et al.* use a recent, faster algorithm[4]. For each vertex, let that vertex be the "center," and calculating the shortest paths to every other vertex from that center. Doing this for a single vertex takes $\Theta(m)$ time. Instead of summing the pair-dependency ratio at the end, the new algorithm accumulates partial pair-dependency ratios along the way, again needing only $\Theta(m)$ time to compute all partial pair-dependencies for a single center. At the end, the accumulated pair-dependency ratios are divided by two, since each path has been counted twice, once from each direction. The computation runs in $\Theta(nm)$ time and requires, for unweighted graphs, $\Theta(n + m)$ space.

To apply the method to the email graph, Tyler *et al.* chose to represent the email exchanges in an unweighed, undirected graph, where an edge exists between two vertices if the corresponding people had, over a period of almost three months, (a) exchanged a total of at least 30 emails, and (b) had sent at least 5 emails in each direction — that is, each person had emailed the other at least 5 times. Only the "To:" and "From:" lines were used. Because some people emailed infrequently or used external systems for email, their original graph of 485 HP Labs employees was reduced to 367 vertices which were connected by 1110 edges.

The threshold for stopping was as follows: a component of l vertices would be identified as a community, output, and removed from the graph when either (a) $l < 6$ or (b) the maximum betweenness of an edge in the component was less than or equal to $l - 1$. The smallest viable community was considered to consist of three vertices, so a component of 5 or fewer vertices could not be split into more than one community. The other condition prevents a leaf vertex from being disconnected. The single edge connecting such a vertex to all $l - 1$ other vertices in a component is on all the shortest paths, so its betweenness is $l - 1$.

When a vertex has an edge to one dense subgraph, an edge to another dense subgraph, and no other edges, both of those edges will have the same betweenness. The choice of which one to remove is arbitrary, and such a vertex "could rightfully be considered to be part of both communities."[26, p86] Furthermore, a single choice of this sort made in early iterations can affect the placement more than one vertex. Because of this, Tyler *et al.* introduce randomness and do an approximate partitioning multiple times, combining the results and identifying the most reliably-appearing communities. For large components, Tyler *et al.* use an algorithm that cycles randomly though j vertices, $j < n$, using each as a center, until a vertex exceeding a certain betweenness is found. The highest-betweenness edge is then removed. When the remaining components are small, the exact algorithm is used. The whole process is repeated i times. The results are aggregated, with vertices placed in a community if they consistently appeared in that community throughout the i runs[28].

Although this removes some edges that are not of maximum betweenness, in experiments, the overall outcome did not substantially differ from that obtained using the exact algorithm.

When run on the email graph, the algorithm found 66 communities with a mean of 8.4 members and standard deviation 5.3. The largest of these was of size 57 and there were several of size 2. (Presumably these either existed in the initial graph as isolated dyads or were determined during aggregation, since no run of the algorithm would ever break a larger graph into communities with fewer than 3 vertices.) Of these 66 communities, 17 crossed formal organizational boundaries. A visualization of the resulting graph showed that the formal leaders of the organization tended to appear near the center.

Tyler *et al.* evaluated the output of their algorithm by interviewing 16 subjects who appeared in their graph in 7 different, arbitrarily-chosen communities. The interviews lasted about 15 minutes and consisted of first asking the subject if the discovered community "made sense," then asking if there were people missing or if people appeared who should not, and finally asking for more general comments about this community. All 16 subjects said that the identified communities were real, saying things like "yes, that's my department," "that's a group that reports to me," and "that's pretty much our project team."[26, p92] Of the 16 interviewees, 7 said that people were missing and 6 said that a person appeared in the community who shouldn't. An intern who was not formally part of a group, but did work with the group, was correctly identified as part of that community. One large community was identified as a department in which there was "a lot of overlap in the projects," even though the interviewee did not personally work with everyone else in the department.[26, p92]

A review of this work co-authored by one of the original researchers suggested that weights on edges could be added to indicate the frequency of communication, and suggested the use of more recent, faster algorithms for community identification.[15]

2.2.2 Static and Temporal Community Structures

Three researchers from Universitè de Genéve and the Weizmann Institute of Science used an information-theoretic approach to email data, and incorporated temporal information via the "Date:" line, to separate static community structures from ones indicated by the co-occurrence of emails.[8] Jean-Pierre Eckmann, Elisha Moses, and Danilo Sergi used 83 days of email data from a university server, culling this to the 309,125 messages that were exchanged between 3,118 internal users. They partitioned the initial graph into a static graph, similar to the one developed by Tyler *et al.*, and also constructed a conjugate graph based on the initial graph, one which incorporated information about how synchronized the email messages between triads were on a day-to-day basis.

Eckmann *et al.* built their static graph in a slightly different way than did Tyler *et al.*, using the concept of a vertex's *curvature.*[7] For a given vertex v with k neighbors, induce the

subgraph $G_\Gamma = (\Gamma_v, E_\Gamma)$ on the neighborhood of v. (Note that the number of edges E_Γ is at most $\binom{k}{2}$.) Now, $curv(v) = (|E_\Gamma|/\binom{k}{2})$. If v and its neighbors form a clique, $curv(v) = 1$, the maximum value. If there is no edge between any of the neighbors of v, $curv(v) = 0$. Considered in the context of small-world networks, curvature is the local, per-vertex analog of clustering. Eckmann *et al.* only included vertices with a curvature of more than 0.1 in their static graph, which revealed "the clear appearance of departmental communities."[8, p4]

Eckmann *et al.* introduced a conjugate graph based on the *temporal cohesion* of triads. These triads, which are vertices in the conjugate graph, correspond to triangles in the original graph. The temporal cohesion of a triangle A, B, C is the mutual information of that triangle for one-day (24-hour) intervals. First, consider the mutual information for a pair of vertices A, B. This measures how much knowledge of the activity of A (sending email, in this case) improves one's guess about the activity of B. It is a symmetric measure. Consider that $p_A(1)$ is the probability that A sends email to B within 24 hours; given the data over d days, this probability is $N_A(1)/d$, the number of days A was observed to send email to B divided by the total number of days. Similarly, $p_A(0) = N_A(0)/d$ expresses the probability that A does not send email to B within 24 hours. Allowing i and j to be binary random variables, the four joint probabilities $p_{AB}(i, j)$ are each defined as $N_{AB}(i, j)/d$. They are as follows: $p_{AB}(0, 0)$, the probability that neither party sends email to the other; $p_{AB}(1, 0)$, the probability that A sends to B but B does not send to A; $p_{AB}(0, 1)$, the probability that B sends to A but A does not send to B; and $p_{AB}(0, 0)$, the probability that the parties both email each other. Now, the mutual information for the pair is as follows:

$$I(A, B) = \sum_{i,j=0,1} p_{AB}(i, j) \cdot log\left(\frac{p_{AB}(i, j)}{p_A(i) \cdot p_B(j)}\right)$$

Consider a pair of people who email each other independently at random, sending emails on 50% of the days, and consider that they email each other on the same day a quarter of the time, the first emails the second without a reply a quarter of the time, and the second emails the first a quarter of the time (as is the case in expectation.) So, for all four probabilities $p_{AB}(i, j)$, the value is .25, and $I(A, B) = 4 \cdot (.25 \cdot log(\frac{.25}{.5 \cdot .5})) = 0$, an appropriate value for uncorrelated A and B. If, on the other hand, A and B email each other on almost half the days and neither emails the other almost half of the days, $I(A, B)$ will be approximately $2 \cdot (.5 \cdot log(\frac{.5}{.5 \cdot .5})) = 1$, consistent with almost complete correlation. If one party emails the other every day, or not at all, the denominator of the log term would be zero, so mutual information is not defined. This is appropriate, since constantly sending email, or never sending email, cannot provide any information about what happens on different days. Presumably, if $p_{AB}(i, j) = 0$ for some i, j, mutual information should be computed by simply omitting the corresponding term from the summation; the log will be $-\infty$, but the mutual information can still have a meaningful value.

Mutual information is defined for triangles analogously. Since six types of email transmissions are possible in a triangle, $p_{ABC}(i_1, i_2, i_3, i_4, i_5, i_6)$ is defined for six binary random variables, and is equal to $N_{ABC}(i_1, i_2, i_3, i_4, i_5, i_6)/d$. Each pair of variables refers to communication

between a possible pairing of users: (i_1, i_2) refers to communication between A and B, with $i_1 = 1$ if A sent a message to B and $i_2 = 1$ if B sent a message to A. Similarly, (i_3, i_4) refers to communication between A and C and (i_5, i_6) to communication between B and C. Then, the temporal cohesion of a triangle is its mutual information:

$$I(A, B, C) = \sum_{i_1..i_6=0,1} p_{ABC}(i_1, i_2, i_3, i_4, i_5, i_6) \cdot log \left(\frac{p_{ABC}(i_1, i_2, i_3, i_4, i_5, i_6)}{p_{AB}(i_1, i_2) \cdot p_{AC}(i_3, i_4) \cdot p_{BC}(i_5, i_6)} \right)$$

In the conjugate graph, any triangle with temporal cohesion $I \geq 0.5$ appears as a node. If two triangles that appear as nodes have a common edge, an edge is present between them in the conjugate graph.

The researchers reported that the conjugate graph includes "many clusters that are new, and do not appear in the high curvature graph. These are typically users that are not in the same department."[8, p4] Two of the clusters represented communities within the university involved in non-academic communications. One cluster represented visiting scientists who were all from the same foreign country. Eckmann *et al.* noted that consideration of email exchanges within companies or government agencies might be more useful than consideration of university email exchanges, since "the major activity in a university is research, which usually involves few individuals."[8, p4]

Critique of Recent Research

Computer-mediated communication and its effect on community has proven a difficult topic, whether researchers have approached the issue from a humanistic perspective or have attempted to deal with it using quantitative techniques and methods from disciplines such as computer science. However, researchers have successfully brought some mathematical and computational insights to bear on the extensive Web and email data that is now available. Worthwhile new techniques for automatically uncovering community structure have resulted. While the work done so far has its limitations, it has advanced our understanding of community as it relates to the Web and email.

3.1 Limitations and Areas for Improvement

One of the main problems with the research surveyed here involves a lack of evaluation of results against a social standard. Researchers tend to present lists or to answer an intrinsic question that refers to characteristics of a graph, leaving open the question of whether the clustering of the graph actually represents some social reality. There are also some possible problems with the underlying models, problems that may be revealed by the heuristics used to shore these models up and improve the results.

3.1.1 Evaluation and the Social World

For the most part, these studies, even the most mathematically impressive ones, stop short of considering whether the "communities" they discover really align with social realities. A pronounced tendency in the current research involves reporting on whether results look right or seem helpful, without trying to establish what social phenomena underly the data and whether or not these real social structures are actually being discovered by the newly-developed techniques.

Community and *friends* are defined in some of these studies purely with reference to the properties of subgraphs. Although a true, extrinsic meaning for these concepts is never established, researchers expect that their discovery of communities and their prediction of whether of not people are friends will correspond to the sociological meanings, or at least to some intuitive meanings, of these terms.

Adamic and Adar did try to informally determine, in some cases, whether the "friends" they identified online were actually friends.[2] But their basic definition of friendship as the existence of any link between home pages was problematic. The structural data worked fairly well to predict friendship, but friendship was, after all, just a name for a structural feature.

Using in-links and out-links from two Web pages to predict the existence of a link between them is unusual. The predicted links would have been part of the original data collected; It is not actually clear whether these links were removed from the data before the experiments were run. Even if these links were removed, there would seem to be few cases in which the value to be predicted (a link between two specific pages) would ever *not* be part of the input data, consisting of all links between all pages crawled.

Gibson *et al.* and Kleinberg present "top authority" lists which appear admirable, as do Flake *et al.*, but these lists are never compared to any others — neither human-generated ones nor ones produced automatically using other methods. While their techniques seem to do well at discovering communities, and the foremost pages identified are certainly plausible as leading sites, Web-conversant experts on those topics could possibly come up with better lists or find problems with the ones that were generated. If lists of top authorities, or community members in general, were built using different techniques, people who were familiar with the Web community could rank or rate the lists, examining the sites in question if necessary to thoroughly evaluate the quality of these lists. Even if the discrepancies between discovered communities were minor, they might reveal biases that would be more strongly expressed in other cases, biases that could then be addressed.

Eckmann *et al.* colored the nodes of their static graph to indicate different departments, showing that different departments fall on different parts of the static graph. But it is not clear how to induce departmental graphs using the static graph they output. In their conjugate graph, they identified three components as corresponding to real social exchanges, but they did not characterize the other components or try to determine how many real communities were not correctly identified by components of the conjugate graph.

3.1.2 Heuristics, Tricks, and Free Parameters

The use of HITS for community identification, by Gibson *et al.* and by Kleinberg, clearly involved the fewest free parameters and the fewest heuristic exceptions to the basic model. Even the problems identified with the system — for instance, sites like Yahoo! and AltaVista being identified as authorities — seem to validate the model rather than argue against it, since the identification of such sites, inappropriate as this identification may be for search purposes, does reflect a true sort of authority that is conferred on them in many different communities online. The claim is made that HITS has only one free parameter, the size of the initial "root set" of pages that is chosen using a search engine and a search term. (As a practical measure, the top 10 hubs and top 10 authorities are usually identified; the choice of how many to list is also a free parameter. Additionally, it would be possible to vary the cap of 50 on the number of pages that will be fetched due to links to a particular page in the root set.) HITS actually inherits the heuristics and free parameters of the search engine that is employed when it starts off with this root set. Whatever information retrieval techniques are used to select this set of pages, and whatever shortcuts and heuristics they employ, essentially become part of HITS. The same cannot be said of the method of Flake

et al.[9], which, although it employs many heuristics, employs only explicit ones, beginning its search with a single URL.

Although Flake *et al.* cite as an advantage of their technique that there is no need to extract multiple eigenvectors,[9, p151] they define community so that each Web page, used as a source, can only be part of a single community. Their technique would be able to identify communities that are in ring-like topologies or otherwise do not have a small number of dominant members, an advantage over HITS. There is a natural way to use HITS to accomplish such identifications, however: look for a prominent cluster of more than 10 hubs and 10 authorities. Community members will still have larger components than non-community members, even if there are not a small number that are dominant. Several heuristics that Flake *et al.* used were justified briefly with reference to the nature of Web communities, but these make the correspondence between the original model and the Web less clear:

1. Four seed vertices were actually used. The additional ones were the vertices that were identified after the previous crawl as being of highest degree. They were connected to the original source with edges of infinite capacity.

2. The first-generation to second-generation edges were made bidirectional. That is, edges were added in the reverse direction between the pages crawled in the second generation and the ones crawled initially, if such edges were not already present.

3. The weight associated with edges between first-generation and second-generation vertices was set at k (a value greater than one), while the weights on all the other edges were left at 1. The proof of correspondence between the original problem and the new minimum cut problem with virtual sink assumes that all the edges in the original graph are given weight k.

4. Only the vertices that are most topologically distant from the source (the ones in the final generation) were connected to the virtual sink. The proof assumes all vertices are connected to the virtual sink.

5. Sometimes only links that cross domain boundaries are added as edges, but in one experiment, edges were added based on links within a domain.

There was no proof of convergence given for the technique as implemented, nor is there a proof of how good an approximation it is to the exact technique.

Adamic and Adar identified and removed links between home pages of users who did not know each other, finding this situation was "easy to detect."[2, p213] The presence of Yahoo! as an authority on a HITS-generated list is consistent with the concept of an authority, at least, but the presence of non-friends who have links between them (the author of a Web page template and someone who used that template, for instance) reveals that equating a "friend" relationship with a home page link is essentially problematic. Also, only student

home pages hosted at Stanford and MIT were used; external ones were not considered, so the problem of identifying a student home page was not tackled in general.

The Tyler *et al.* technique involved first creating an unweighed graph, using a threshold of 30 messages (with at least 5 sent in each direction) to determine the presence of an edge. As the researchers note,[26, pp94–95] this skews results in favor of people who send more email messages. Using a weighted graph and normalizing each user's messages over the total number they sent would have added complexity but may have removed the need for this threshold parameter. Messages sent to a list of more than 10 recipients were also excluded; this cutoff might have been changed so that messages sent to larger groups were smoothly discounted, rather than being cut off sharply at 10.

The randomized procedure used by Tyler *et al.* cycled through j vertices on each iteration and was repeated i times, providing two more free parameters. There are ways to let a single vertex lie in multiple communities without employing this sort of randomization. For instance, if there are two edges x and y with maximum betweenness, deterministically replicate the whole graph and delete x in the first copy and y in the second. Then, continue the edge deletion process on both graphs. Clearly, this increases the time and space complexity of the operation, but exact techniques are worth studying in further detail so that the quality of approximations can be demonstrated, and in case there are fast formulations of these exact techniques to be discovered.

Eckmann *et al.* developed a technique that involved few heuristic adjustments and free parameters. They used a cutoff of 24 hours to determine temporal cohesion — a meaningful time unit, although others could be used, and different types of cohesion might be found over different intervals. It also may be possible to fruitfully omit the range of time over which a person almost never sends email (because he or she is sleeping, for instance) from the interval considered, so that an email response sent first thing in the morning is counted as occurring right after an email in the middle of the night. A cutoff of 0.1 curvature was used to determine if a node would appear in the static graph; a triangle's mutual information needed to be at least 0.5 for the corresponding node to appear in the temporal graph. It is possible that these cutoff points, and the appropriate way to deal with the time interval, need to be found on a per-organization basis. A principled way for finding the appropriate values in general may need to be developed.

3.1.3 The Fringe of the Email Graph

Both email studies discarded messages that had "To:" or "From:" lines from outside the organization being studied. Eckmann *et al.* wrote that this was done "since external links are necessarily incomplete."[8, p1] They are incomplete, although all the information pertaining to the organization being studied is complete. That is, if *someone@elsewhere.org* is an external email address, all email to *someone@elsewhere.org* that is from anyone within the organization being studied, and all email from *someone@elsewhere.org* to internal addresses,

is logged by the organization's server. It is possible that *someone@elsewhere.org* would make the difference between a set of vertices being considered a community or not. This email address may represent a collaborator, granting agency contact, vendor, or other individual who is important to the organization. It is worth some effort to incorporate such external data into an analysis. However, email correspondence between two external addresses in the data set will not be logged by the organization's server, so external data is indeed in a different category — its partial representation gives these vertices an artificially low degree, for instance. Simply including it as if it were internal would probably not suffice. One approach is suggested by work on call graphs that adds "pseudo edges" to reflect local phone calls that are not reflected in the data collected by a long-distance provider.[6]

3.2 Community Polarization and HITS

A feature of HITS has been used for separating a polarized community. The use of this feature to characterize a community as polarized, or to quickly determine the structure of a community, can be misleading, and is worth some discussion. (Here, *polarized* is used to mean "concentrated around two conflicting positions.") Kleinberg gives a result in which the community of abortion-related Web pages has been divided along a "natural separation" into pro-choice and pro-life pages just with reference to the signs of the second leading eigenvector's components.[18, p625] The result has been influential, and other researchers have since tried to characterize pro-choice and pro-life Web pages.[1]

Kleinberg does not explicitly state that the signs of an eigenvector's components can be used to detect or analyze polarized communities, only noting that the abortion Web community he examined is a polarized community and that work in spectral heuristics for partitioning undirected graphs has found that vertices associated with components of different signs "are often well-separated."[18, p623] However, the strong suggestion of this result — that large positive components are typically associated with one pole of a community and large negative components are associated with that community's opposite pole — is potentially misleading. Looking at the signs of components can help determine what authorities are on either end of a community already known to be polarized, but it can also suggest that certain communities are polarized when they are not. If the set A corresponds to large positive components and the set B to large negative components, the essential observation made in the spectral partitioning of graphs [18, 5] was only that the elements in A tend to be well-separated from the ones in B. It is not always observed that A and B are sets of nearby vertices on the opposite ends of a single pole.

Consider a graph representing Web pages in a console gaming community. A non-principal eigenvector corresponding to authorities in this community might have somewhat isolated sets P (Playstation 2), X (Xbox), G (Gamecube), and D (Dreamcast), all of which are linked to by common hubs but which tend to not refer to one another. It may happen that P and G have large positive components and X and D large negative components. But the signs

of the eigenvector's components cannot signal that P and G are also separated from one another, or that X and D are also well-separated. After slight changes in the graph, of the sort that might be observed after the search engine has recrawled, it may happen that P and D end up on one "pole" and X and G on the other. If we interpreted the positive-valued and negative-valued top authorities to be polarized pages, it would look as if something radical had happened to the community, which would not have been the case.

Thus, using the sign of the components to characterize the top authorities can suggest a binary separation when a community is not actually polarized. The presence of large positive components and large negative components only signal that there are some important nodes that are well-separated. An analysis of the identified authorities (in terms of their topological distance from one another) will still be necessary to determine what the structure of the main authorities is like. Since the difference in sign does not apply, anyway, in the case of the principal eigenvector, which has all positive components, it is always necessary to analyze the top authorities in the principal community using another technique if anything is to be known about their structure. Although this matter may seem minor, failing to comment on it would risk letting a computer science result reinforce a common rhetorical and journalistic fallacy, that all issues have exactly two sides to them and that one only needs to identify proponents and opponents.

3.3 Significant Advances

Much of the research considered here has made evident the importance of communications structure, even when it is considered apart from the content of communications, and has shown that this structure is susceptible to automated analysis. This work has also brought sophisticated mathematical and algorithmic techniques to bear, and has motivated advances in algorithms for large, sparse social graphs. Finally, one group of researchers has tried to connect their work on community discovery to social reality by evaluating it with interviews and using social relationships, rather than any intrinsic feature of the data, as their standard.

3.3.1 The Importance of Structure

Kleinberg's work on HITS, the basis of the investigation of community by Gibson *et al.*, revealed an important aspect of the Web, one that seems obvious in retrospect but had not been fully appreciated: a Web page can be considered good at pointing to other pages, or it can be good at being pointed to by other pages. The idea of hubs and authorities was not developed in bibliometrics, perhaps, as Kleinberg speculates[18], because treating every document as an "authority" was good enough when a body of academic articles was being considered. The Web is a different sort of network, structurally and in terms of the types of material online. Web pages can point to each other in a way that does not have

a clear analog in the academic literature — although one article might cite a forthcoming article that cites the original article, this situation is far less natural that the one of Web pages pointing to one another. Research on Web communities has not just built upon the work done in bibliometrics, but has revised this work significantly for the new context as Web-native techniques have been developed.

The research done on finding "friends" by considering student home pages and mailing list memberships[2] was one attempt to consider the usefulness of structural data as opposed to textual content. It highlighted that structure can often speak as strongly as or more strongly than the text. The results of that study were also interesting for exposing some qualitative differences between the Stanford and MIT student communities.

Analysis of email communities, similarly, has gone beyond mere application of the work done on the structure of the Web. Researchers have identified the relationship between the email graph and the social network of acquaintanceship and collaboration[26] and have noted differences between the nature of email and Web graphs.[8] The co-occurrence of emails in time has been shown to be useful in creating a conjugate graph that reveals additional sorts of social structures.[8]

3.3.2 Mathematical, Algorithmic, and Social Network Techniques

Research has brought together insights from linear algebra, graph algorithms, and social networking to create techniques that are better suited for the Web and email than generic social network analysis methods or bibliometric techniques would be. Some progress on graph algorithms, such as that made by Brandes[4], has been specifically motivated by social network analysis measures and the need to apply these to large, sparse graphs, of the sort found in Web and email data. Researchers working directly on community discovery have also made advances.

The notion of hubs and authorities, introduced by Kleinberg and used in community analysis by Gibson *et al.*, is a powerful one that models important aspects of the Web. By showing the correspondence between the iterative algorithm for hubs and authorities and the determination of the leading eigenvectors of two matrices, progress on two fronts was made possible. First, this mapping was convenient in terms of time complexity and allowed HITS to work more quickly by exploiting fast existing algorithms for determining the leading eigenvector. Second, the reformulation suggested a way to identify non-principal communities by solving for other eigenvectors. This leap would have been difficult to make had the iterative algorithm not been formulated as a matrix problem.

Eckmann *et al.* made a valuable distinction between temporally cohesive edges and those that are significant in a "static" graph (which is actually a sort of temporal graph that considers all communications in a single, large time-slice). This pointed to a new dimension of structure that can be used in the analysis of communities. By further refining our un-

derstanding of temporal data, teams of people who are involved in short-term projects, but whose communications are not cohesive after these projects end, could also be identified.

3.3.3 Connecting Graphs to People

One group of researchers, Tyler *et al.*, took the important step of performing a formal and extrinsic evaluation of their results, interviewing people who were identified as being in particular communities and asking them if the identification was plausible. To determine the relationship of online communications with online and offline communities, it is particularly important to survey or interview community members, who are good sources of information about the existence of communities. Conducting interviews can help in detecting shortcomings or pathologies of automatic community discovery techniques. It may also reveal the inherent limitations of using information about the structure of email or Web communications. While structural information has been surprisingly rich and has proven useful for community discovery, we can only evaluate how effective techniques that are based upon it are if we look beyond the graphs we create to the social relationships we are attempting to model.

Acknowledgments

This report was written for my Written Preliminary Exam II. I appreciate the help of my WPE-II committee: Fernando Pereira, the chair, whose detailed comments on the report were particularly useful; Sampath Kannan; and my advisor, Michael Kearns. Thanks also to Ryan McDonald and Hanna Wallach for reading through a draft of the report and pointing out typographical and other errors.

Bibliography

[1] Lada A. Adamic. The small world web. In *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries*, pages 443–452. Springer-Verlag, 1999.

[2] Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[3] Stephen P. Borgatti and Pacy C. Foster. The network paradigm in organizational research: a review and typology. *Journal of Management*, 29(6):991–1013, 2003.

[4] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[5] Fan R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, Providence, R.I., 1997.

[6] Corinna Cortes, Daryl Pregibon, and Chris Volinsky. Computational methods for dynamic graphs. *Journal of Computational & Graphical Statistics*, 12(4):950–970, 1 December 2003.

[7] Jean-Pierre Eckmann and Elisha Moses. Curvature of co-links uncovers hidden thematic layers in the world wide web. *Proceedings of the National Academy of Science USA*, 99:5825–5829, 2002.

[8] Jean-Pierre Eckmann, Elisha Moses, and Danilo Sergi. Dialog in e-mail traffic. `http://xyz.lanl.gov/abs/cond-mat/0304433`, 18 April 2003.

[9] Gary William Flake, Steve Lawrence, and C. Lee Giles. Efficient identification of web communities. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–160. ACM Press, 2000.

[10] David Gibson, Jon Kleinberg, and Prabhakar Raghavan. Inferring web communities from link topology. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia*, pages 225–234. ACM Press, 1998.

[11] Michele Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Science USA*, 99:8271–8276, 2002.

[12] Malcolm Gladwell. *The tipping point: how little things can make a big difference*. Little, Brown and Company, 2000.

[13] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1989.

[14] George Hillery. Definitions of community: areas of agreement. *Rural Sociology*, 20(2):111–123, 1955.

[15] Bernardo A. Huberman and Lada A. Adamic. Information dynamics in the networked world. In Eli Ben-Naim, Hans Frauenfelder, and Zoltan Toroczkai, editors, *Complex Networks*. Springer, 2004.

[16] Hitoshi Isahara and Hiromi Ozaku. Intelligent network news reader. In *Proceedings of the 2nd international conference on intelligent user interfaces*, pages 237–240. ACM Press, 1997.

[17] Faisal M. Khan, Todd A. Fisher, Lori Shuler, Tianhao Wu, and William M. Pottenger. Mining chat-room conversations for social and semantic interactions. `http:// www.lehigh.edu/images/userImages/jgs2/Page_3471/LU-CSE-02-011.pdf`, 2002.

[18] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, September 1999.

[19] J. C. R. Licklider. Topics for discussion at forthcoming meeting. MIT Institute Archives. Memorandum, 23 April 1963.

[20] J. C. R. Licklider. In memoriam: J. C. R. Licklider 1915-1990. Reprint of Man-computer symbiosis (1960) and The computer as a communication device (1968) `http://gatekeeper.research.compaq.com/pub/DEC/SRC/research-reports/abstracts/src-rr-061.html`, 1990.

[21] William G. Ouchi. Markets, bureaucracies, and clans. *Administrative Science Quarterly*, 25(1):129–141, March 1980.

[22] Howard Rheingold. *The virtual community: homesteading on the electronic frontier*. Addison Wesley, Reading, Mass., 1993.

[23] Warren Sack. Conversation map: an interface for very large-scale conversations. *Journal of Management Information Systems*, 3(3):73–92, Winter 2001.

[24] John Scott. *Social network analysis: a handbook*. SAGE Publications, Thousand Oaks, Calif., 2nd. edition, 2000.

[25] Marc Smith. Invisible crowds in cyberspace: measuring and mapping the social structure of USENET. In Marc Smith and Peter Kollock, editors, *Communities in cyberspace*, pages 195–219. Routledge, 1999.

[26] Joshua R. Tyler, Dennis M. Wilkinson, and Bernardo A. Huberman. Email as spectroscopy: automated discovery of community structure within organizations. In *Proceedings of the International Conference on Communities and Technologies*, pages 81–96. Kluwer Academic Publishers, 2003.

[27] Barry Wellman. Computer networks as social networks. *Science*, 293:2031–2034, 14 September 2001.

[28] Dennis M. Wilkinson and Bernando A. Huberman. A method for finding communities of related genes. *Proceedings of the National Academy of Science USA*, 101(Suppl. 1):5241–5258, 6 April 2004.