

Discovering Shared Interests Among People Using Graph Analysis of Global Electronic Mail Traffic

Michael F. Schwartz

David C. M. Wood

Department of Computer Science
University of Colorado
Boulder, Colorado 80309-0430
+1 303 492 7514
{schwartz,dcmwood}@cs.colorado.edu

October 1992

To appear, *Communications of the Association for Computing Machinery*

Abstract

An important problem faced by users of large networks is how to discover resources of interest, such as data or people. In this paper we focus on locating people with particular interests or expertise. The usual approach is to build interest group lists from explicitly registered data. However, doing so assumes one knows what lists should be built, and who should be included in each list. We present an alternative approach, which can support a more fine grained and dynamically adaptive notion of shared interests. Our approach deduces interests from the history of electronic mail communication, using a set of heuristic graph algorithms. We demonstrate the algorithms by applying them to data collected from 15 sites for two months. Using these algorithms we were able to deduce shared interest lists for people far beyond the data collection sites. The algorithms we present are powerful, and if abused can threaten privacy. We propose guidelines that we believe should underly the ethical use of these algorithms. We discuss several possible applications that we believe do not threaten privacy, including discovering resources other than people, such as file system data.

CR Categories and Subject Descriptors: C.2.4 [Computer-Communication Networks]: Distributed Systems – *distributed applications*; G.2.2 [Discrete Mathematics]: Graph Theory – *graph algorithms*; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – *abstracting methods*; H.3.4 [Information Storage and Retrieval]: Systems and Software – *information networks*; H.4.3 [Information Systems Applications]: Communications Applications – *bulletin boards, electronic mail*; H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces – *organizational design*; I.5.2 [Pattern Recognition]: Design Methodology – *classifier design and evaluation*; I.5.3 [Pattern Recognition]: Clustering – *algorithms, similarity measures*; I.5.4 [Pattern Recognition]: Applications; K.4.1 [Computers And Society]: Public Policy Issues – *privacy*.

General Terms: Design, Experimentation, Measurement

Additional Key Words and Phrases: resource discovery, directory service, dynamic organization, interpersonal communications, collaboration, graph theory, traffic analysis

1. Introduction

Ongoing increases in wide-area network connectivity promise vastly increased opportunities for collaboration and resource sharing. A fundamental problem that confronts users of such networks is how to discover the existence of resources of interest, such as files, retail products, network services, or people. In this paper we focus on the problem of discovering people who have particular interests or expertise. For an overview of the larger research project into which this work fits, the reader is referred to [Schwartz 1992].

The typical approach to locating people is to build a directory from explicitly registered data. This approach is taken, for example, by the X.500 directory service standard [CCITT/ISO 1988]. While this approach provides good support for locating particular users (the "white pages" problem), it does not easily support finding users who have particular interests or expertise (the "yellow pages" problem). One could create special interest group lists, but doing so requires a significant amount of effort. For each group someone has to build and maintain a membership list. Moreover, building such lists assumes one knows which lists should be compiled, and who should be included in each list. In a large network, the set of possible interest groups can be quite large and rapidly evolving. It is difficult to track the interests of such a community using explicitly registered data.

We consider a different approach, which deduces shared interest relationships between people based on the history of electronic mail communication. Using this approach, a user could search for people by requesting a list of people whose interests are similar to several people known to have the interest in question. This technique can support a fine grained, dynamic means of locating people with related interests. The set of possible interests can be arbitrarily specialized, and the people located will be appropriate at the time of the search, rather than at some earlier time when a list was compiled.

One might attempt to discern shared interests by analyzing subject lines and message bodies in electronic mail messages. Beyond the obvious privacy problems, doing this would pose difficult natural language recognition problems. Instead, we approached the problem by analyzing the structure of the graph formed from "From:/To:" electronic mail logs, using a set of heuristic graph algorithms. We demonstrate the algorithms by applying them to electronic mail logs we collected from 15 sites around the world between December 1, 1988 and January 31, 1989. The graph generated from these logs contained approximately 50,000 people in 3,700 different sites world-wide. Using these algorithms we were able to deduce shared interest lists for people far beyond the data collection sites.

Because the algorithms we present can deduce shared interest relationships from any communication graph, they are powerful and can threaten privacy. We propose recommendations that we believe should underly the ethical use of these algorithms, and discuss several possible applications that we believe do not threaten privacy.

We begin in Section 2 by defining a model of the shared interest relationships that we wish to analyze. In Section 3 we discuss the methods used to collect and filter the data. In Section 4 we discuss the scope of the sampled data, and how we eliminated data that interfered with the interest clustering algorithms. In Section 5 we present the interest clustering algorithms. In Section 6 we propose guidelines for the ethical use of these algorithms, and suggest applications that fit these guidelines. We also discuss possibilities for adapting the algorithms to other types of resources besides people, such as file system data. In Section 7 we discuss related work. Finally, in Section 8 we offer our conclusions.

2. Specialization Subgraph Structure

When people collaborate, they often form "networks" of colleagues according to shared interests or responsibilities, both within and across bureaucratic boundaries. To model this organizational mechanism, we define a clustering relationship called a *Specialization Subgraph (SSG)*. An SSG of a communication graph is a subset of nodes and edges, where each of the people (nodes) share a common interest, which is reflected in the individual communications (edges). While simply communicating with another person does not imply shared interests, being in an SSG with that person does imply shared interests. In general, a person will belong to many different SSGs according to her or his various interests and responsibilities, and can try to locate information about a particular topic or accomplish a specific task by consulting an appropriate SSG. The goal of the algorithms we present in this paper is to derive SSGs for particular interests from a communication graph.

As an example of how people collaborate through SSGs, one of the current paper's authors (Schwartz) became interested in finding measurements of the number of naming domains in the Internet¹ [Mockapetris 1987], to help

¹ Throughout this paper we use "internet" to refer to general internetworks. We use "Internet" to refer specifically to the growing collection of

estimate the scope of an Internet directory facility that he built [Schwartz & Tsirigotis 1991]. To locate such measurements, Schwartz sent electronic mail messages to a small number of people in a relevant SSG, the members of which are shown in Table 1.

Person	Affiliation at Time of Study	Relevant Interests at Time of Study
Phil Karn	Bell Communications Research	Networking Researcher
Sol Lederman	SRI Network Information Center	Network Information Center Staff
Mark Lottor	SRI Network Information Center	Performed Measurements of Interest
Paul Mockapetris	USC Information Sciences Institute	Designer of Domain Naming System
Michael Patton	MIT Laboratory for Computer Science	Network Manager
Larry Peterson	University of Arizona	Naming & Networking Researcher
Marshall Rose	Performance Systems International	Manager, White Pages Pilot Project
Karen Sollins	MIT Laboratory for Computer Science	Naming & Networking Researcher

Table 1: Specialization Subgraph Used in Example Resource Discovery Process

The referral graph is illustrated in Figure 1. In this figure, a directed edge from node x to node y indicates that x referred Schwartz to y . Some people suggested that Schwartz send messages to particular mailing lists or electronic bulletin boards. Schwartz did not do so, since these groups represent much larger and less focused SSGs. As can be seen, the search quickly converged on Lottor, who had performed the measurements of interest. Note also that closeness of interest did not always correspond to geographical or organizational proximity. For example, although Lederman was at the same institution as Lottor, he did not know about Lottor, while people in organizations 2,500 miles away did. Lottor has since published his study results [Lottor 1992].

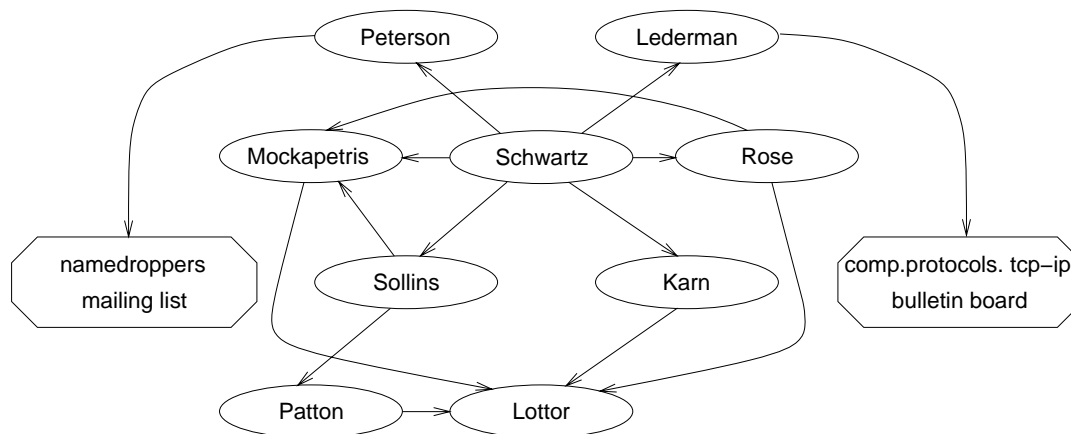


Figure 1: Referral Graph for Example Resource Discovery Process

3. Data Collection and Transformation

To explore algorithms for locating SSGs within a communication graph, we collected data using a pair of program scripts. The first script ran at each participating site, collecting and mailing the data to the second script, which ran on a machine at the University of Colorado. The sending script ran once per day for the duration of the study, collecting "From:" and "To:" lines recorded in the UNIX² "syslog" log files for each message transferred by

interconnected IP [Postel 1981] networks that join academic, industrial, and government institutions world-wide. We use "Matrix" to include the global collection of systems interconnected by electronic mail [Quarterman 1990], which extends far beyond the Internet by including many dial-up systems.

² UNIX is a registered trademark of UNIX System Laboratories.

"sendmail", the Berkeley UNIX mail transfer agent [Allman 1985]. This script filtered irrelevant log entries, enforced an upper bound on the number of bytes sent, and fragmented the data into pieces small enough to fit through electronic mail. The receiving script made some basic data validity checks, and saved the data according to the sending site, date, and message sequence number.

Four steps were needed to transform the collected data into a form that could be analyzed. Because a wide variety of mail naming conventions exists, the first step was to transform graph node names into a single canonical format. This allowed us to merge aliases in a number of cases (e.g., "user1@host1.domain" and "host1.domain!user1"). The second step was to discard nodes and edges caused by erroneously specified mail names, or by misconfigured message transfer agents that transformed names incorrectly when passing messages to subsidiary agents. Invalid nodes were detected syntactically (e.g., node names containing multiple "@" signs) as well as from delivery failure indications in the log files. The third step was to remove nodes representing generic functions not associated with shared interests (such as "operator"). We did not discard messages addressed to mailing lists concerning discernible shared interests (such as "performance@cs.wisc.edu"), since such correspondence represents true shared interests, even if the sender of a message does not personally know each of the recipients. Approximately 50 special cases of discernibly invalid interests were checked, derived from manual inspection of the data. The final step was to generate a numeric representation of the graph, for computational efficiency. For example, the graph (user1@host1.domain1—user2@host2.domain2, user2@host2.domain2—user3@host3.domain3) would be transformed to (1—2, 2—3), plus a table listing the transformations, to allow the results of graph computations to be mapped back into symbolic form. This process generated an undirected graph, because the directions of electronic mail transmission turn out not to be important for the clustering algorithms introduced in Section 5.

In total, we collected logs concerning 1,234,862 mail messages, and removed 203,636 (16.49%) of them during the data transformation stages. An interesting peripheral fact is that 130,275 (63.97%) of the invalid messages were not unique, indicating that a large amount of global processing effort was expended trying multiple times to deliver undeliverable messages.

In some cases the data transformations may have discarded valid data. For example, some system administrators send mail concerning true shared interests from the "operator" login. However, discarding data simply reduced the size of what was already a sampled subgraph of the global electronic mail communication graph. We consider the sample size in Sections 4 and 5.5.

4. Sample Scope and Core Isolation

In soliciting study participants, we promised prospective sites that we would not reveal any identifying details of the data we collected. Even so, several sites declined participation, citing privacy concerns. Our ability to obtain study participants was also complicated by the Internet worm of November 1988, which invaded thousands of Internet sites and raised the security consciousness and workloads of most system administrators shortly before we started collecting data [Spafford 1989].

We asked 62 different sites to participate. Among these, fifteen sites containing 22 machines that logged mail traffic (hereafter referred to as "log hosts") did so. Some domains transmitted data from multiple log hosts, corresponding to separately administered computing facilities (referred to from here on as "administrative domains"). The participating administrative domains spanned a range of different characteristics:

- 5 were on the U.S. West Coast
- 2 were in the U.S. Mountain Region
- 4 were in the U.S. Central Region
- 2 were on the U.S. East Coast
- 2 were in Western Europe
- 11 were universities
- 3 were research laboratories
- 1 was a product development firm
- 4 were small (50-200 mail users)
- 7 were medium sized (200-1,000 mail users)
- 4 were large (1,000+ mail users)

While these administrative domains were a small fraction of the domains in the world, the collected data provided information about a surprisingly far reaching segment of the global electronic mail Matrix. The logs recorded information about messages passing between 50,834 users on 17,312 hosts in 3,739 administrative domains, distributed among the following 31 countries:

- Argentina
- Australia
- Austria
- Belgium
- Brazil
- Canada
- Chile
- Denmark
- Finland
- France
- Germany
- Greece
- Iceland
- Ireland
- Israel
- Italy
- Japan
- Malaysia
- Netherlands
- New Zealand
- Norway
- Panama
- Portugal
- Republic of China
- S. Korea
- Singapore
- Spain
- Sweden
- Switzerland
- United Kingdom
- United States

The distribution of sites is illustrated in Figure 2, for host names that we were able to translate into longitude/latitude information (approximately 50% of the hosts). The U.S. was particularly heavily represented in this map because 13 of the 15 data collection sites were located in the U.S., and because the U.S. had more networked sites than other countries in the world at the time data was collected. The map also shows areas of concentration significantly beyond our data collection sites. For instance, we did not collect any data in Boston, yet Boston is heavily represented in Figure 2.

Figure 2: Scope of Captured Electronic Mail Interactions (Mappable Subset)

After only a few days of data collection, the number of new unique edges collected per day decreased dramatically, to approximately 10% of the initial days' collections. This indicates that our data collection captured most of the nodes that might be seen over a longer period.

4.1. Core Isolation

Because of the sampled nature of the data, parts of the graph contained too little information to be analyzed effectively. In this section we present the series of steps we performed to reduce the graph to a *core* subset for which shared interests could be effectively deduced. These transformations will be used in Section 5.

The first step in reducing the graph was observing that it contained 2,557 disjoint components. One component (which we refer to as the *main* component) contained most of the nodes and edges from the full graph. The second largest component contained only 28 nodes and 31 edges, and 1,664 (65.1%) of the components consisted of single edges. These measurements are summarized in Table 2. Throughout the remainder of this paper, we refer to the main graph component as simply "the graph."

The next step in isolating a core subgraph was to measure the graph diameter.³ This step was motivated by the

³ The diameter of a graph is the number of edges in the longest path between any two nodes among a set of shortest paths between all nodes.

	Full Graph	Main Component	Second Largest Component	Number of Single Edge Components
Nodes	50,834	43,498	28	3,328
Edges	183,833	179,030	31	1,664

Table 2: Graph Components

so-called "small world" phenomenon, which indicates that the diameter of a large human social network is typically quite small [Travers & Milgram 1969]. A game based on the co-author relationship with the prolific mathematician Paul Erdős provides a small example of this phenomenon. In this game, a person's *Erdős number* is defined to be one if that person has written a paper with Erdős, two if they have written a paper with someone who has written a paper with Erdős, etc. Based on this definition, the highest Erdős number known to be possessed by a person is seven [Hoffman 1987]. Lore has it that the diameter of the entire world may be as small as six or seven.

Our intent in measuring the diameter of the electronic mail graph was to locate a subgraph that exhibited the small world phenomenon, since that subgraph was probably well enough represented in the data sample that it could support effective shared interest analysis.

Using breadth first search, we found that the electronic mail graph diameter was 21. While small relative to the graph size, this value is larger than suggested by the small world phenomenon. To see if a subset of the graph had smaller diameter, we measured the average and maximum path lengths from each of a randomly selected set of nodes to all nodes in the graph. We found that indeed these measures were significantly smaller than the diameter, as summarized in Table 3. (The accuracy figures given are for 95% confidence intervals.)

Diameter	Sampled Average Path Length (accuracy)	Sampled Maximum Path Length (accuracy)
21	5.96 ($\pm .14$)	13.99 ($\pm .15$)

Table 3: Diameter and Sampled Path Lengths of Main Graph Component

We hypothesized that the graph contained a sizable subgraph with a small diameter, surrounded by a set of nodes for which our data sample captured only sparse interconnections. To isolate the core, we iteratively constructed a sequence of subgraphs, each obtained by removing all of the nodes with degree 1 and their incident edges from the previous subgraph, until no such nodes remained.⁴ The algorithm reduced the graph as charted in Table 4. Iteration 0 was the original graph.

Iteration Number	Nodes	Edges	Iteration Number	Nodes	Edges
0	43,498	179,030	4	20,084	155,616
1	21,689	157,221	5	20,058	155,590
2	20,463	155,995	6	20,051	155,583
3	20,161	155,693	7	20,050	155,582

Table 4: Main Component Core Subgraph Isolation Results

The number of iterations executed roughly corroborated our hypothesis, hinting that there was a core subgraph surrounded by a loose, web-like outer structure 7 links deep (which could add up to 14 edges to the core diameter). The fact that most of the outer structure was removed during the first iteration indicates that there was a sharp

⁴ The degree of a node is the number of edges connected to that node.

distinction between the core subgraph and the outer structure, and hence that the members of the outer structure were probably peripheral to most of the communication taking place in the core subgraph. Another indication of this fact is that at the end of iteration 7 only 46% of the nodes in the full graph remained, yet these nodes were connected to 87% of the edges. A final indication is that removing the outer web caused the average node degree to jump from 7.07 to 45.41. All of these measurements indicate that the sampled data contained relatively little information about the communication structure of the outer web. It makes sense to remove this web when performing interest clustering analysis.

As a visual representation of the results of the core subgraph isolation, Figure 3a shows a map of the U.S. portion of the collected mail edges, while Figure 3b shows the isolated core of Figure 3a. The core map appears to have "hubs" at many of the sites where we collected data (but with edges connecting to many more sites than just where data was collected), underscoring the appropriateness of this graph reduction.

a. U.S. Portion of Collected Mail Edges

b. Core Subgraph of Figure 3a

Figure 3: Visualization of Core Subgraph Isolation

5. Interest Clustering Algorithms

The electronic mail graph contained edges corresponding to many different sets of shared interests. The algorithm we present in Section 5.1 clusters people by shared interest based on the notion of an *Aggregate Specialization Graph (ASG)*. An ASG is a set of Specialization Subgraphs (SSGs) concerning topics of high interest to that person. For example, Schwartz belongs to an ASG that includes interests in networks, distributed systems, privacy/security, and various recreational interests.

An ASG does not directly indicate interest relationships between people, since it concerns several interests, and one cannot tell which subset of the interests placed particular people in the ASG. In Section 5.3 we present an algorithm that uses the ASG isolation algorithm iteratively to find people who share a particular interest, by starting with a specified set of people known to have that interest. We consider implications and possible applications of this latter algorithm in Section 6.

The algorithms we present here are heuristic in nature. If the ASG algorithm works poorly, it may include some people whose interests are not close to the person around which the ASG is computed, or it may fail to include some people whose interests are close to that person's. The goal of applying the algorithm to individuals using the global electronic mail data we collected is to indicate how well the algorithms work. We discuss the heuristic nature of the

algorithms in Section 5.4.

5.1. Aggregate Specialization Graph Isolation Algorithm

Intuitively, we want an algorithm that can isolate ASGs within the graph by searching for highly interconnected subsets of nodes. Searching for a globally maximum subset is infeasible, since that is equivalent to searching for a maximum graph clique, which is NP-complete [Sedgewick 1988]. Instead, we chose an approach whereby a highly interconnected subgraph was constructed around a particular distinguished node. At the heart of this approach is the need for a quantitative formulation of the interest distance between two nodes. Given such a measure, nodes can be sorted in increasing order of distance from the distinguished node.

Our algorithm design method involved proposing a measure of interest distance, running the algorithm using one of the paper's authors (Schwartz) as the distinguished node, and then studying the resulting list to see if it made sense, in terms of the interests known to be possessed by the people in the list. We also studied the results of applying the various algorithms to approximately 40 other people whose interests we knew, to understand ways in which the computations clustered people erroneously.

The basic interest distance measure we used computed the proportion of neighbors that two nodes n_1 and n_2 do not have in common (the symmetric difference set), out of the set of all neighbors of both nodes:

$$InterestDistance(n_1, n_2) = \frac{|(C(n_1) \cup C(n_2)) - (C(n_1) \cap C(n_2))|}{|C(n_1) \cup C(n_2)|},$$

where $C(n_i)$ is the set of nodes directly connected to node n_i . This measure ranges from 0 for two nodes that have all neighbors in common with one another to 1 for two nodes that have no neighbors in common, and decreases as the number of nodes outside the intersection set increases. Using this measure we computed an ASG surrounding the distinguished node as the portion of the computed node list with distance less than 1.0. Intuitively, this measure generates ASGs by locating collections of people whose principal SSGs overlap, as indicated by having few edges outside of these SSGs present in the communication graph.

Applying this algorithm with Schwartz as the distinguished node, we noticed that the results were muddled by the presence of irrelevant people involved in local administrative mail. This problem made it appear that the people most closely related to Schwartz were members of his local department, because there were particularly many interactions with secretarial and computer systems administrative staff, effectively forming pronounced "bridges" between everyone in the department. This type of communication typically does not represent collaborative interests, and should be excluded from the ASG lists.

If enough information were available, one could avoid this problem by building a list of nodes whose roles are not central to the SSGs of interest (such as systems administrators and secretaries), and excluding these nodes and their incident edges. However, given our goal of extracting shared interest information solely from the communication graph, we sought a different solution.

Since the misleading edges derived primarily from intra-organizational communications, we tried performing the interest clustering computations on a derived graph that eliminated intra-organizational communication in two stages. First, we derived a graph where nodes were of the form "user@domain", without the leading host name component. This step eliminates the problem of users having accounts on many workstations within a domain, at the cost of making the (somewhat optimistic) assumption that user names are unique within an administrative domain. Second, we derived a subgraph by including only inter-administrative domain edges (e.g., "user1@cs.colorado.edu—user2@lcs.mit.edu"). Clearly, this graph can still contain edges between people who do not share interests. Yet, as discussed below, the graph contains fewer such edges. Note also that removing intra-domain edges does not mean that two people within a domain cannot be placed in the same ASG. To be placed in the same ASG, two people need only be aggregated with other people from outside domains. Intuitively, this method of aggregating people within a domain is analogous to saying that two professors from the same department share interests if they both attend a particular conference.

Applying the ASG isolation algorithm using the derived inter-domain graph yielded better results, providing a list that more closely corresponded to people whose primary interests matched Schwartz's. However, the list still contained many erroneous entries, for people whom we knew did not have interests related to Schwartz's. Upon further examination we found the problem was that nodes with small degrees were more likely to be considered

close to each other if they shared even a small number of common neighbors, because the size of their symmetric difference set was small. To remedy this problem, we used the algorithm presented in Section 4.1 to derive a core subgraph of the inter-domain graph. This algorithm removed nodes about which too little information was known (because of the sampled nature of the graph), leaving a subgraph with 3,802 nodes and 9,568 edges.

Applying the interest clustering computation to this subgraph achieved dramatically improved results. As an example, the first 12 entries of Schwartz’s computed ASG are shown in Table 5. As can be seen, in most cases the people isolated by this algorithm had interests related to Schwartz’s principal interests. The people in this list were not trivially derivable from the communication graph. Several of the people were not known personally by Schwartz, indicating that the algorithm uncovered relevant people with whom Schwartz had never exchanged electronic mail. In fact, Table 5 includes one person who Schwartz did not know at the time of this study, to whom Schwartz was later introduced by a third person, because their work was related. Moreover, the list omits a number of people with whom Schwartz directly communicated, who do not belong in Schwartz’s ASG. Finally, many of the people in the table were not located at any of our data collection sites. The algorithm was able to derive information about these people given data about only a tiny proportion of the total electronic mail community (15 sites around the world).

Relationship to Schwartz	Distance from Schwartz
Schwartz	0.00
Graduate student where Schwartz attended graduate school, studying networking	0.77
Theory professor who attended graduate school with Schwartz	0.85
Unknown student at a Southwestern U.S. university	0.86
Industrial systems researcher who attended graduate school with Schwartz	0.88
Schwartz’s Ph.D. advisor (interested in performance and distributed systems)	0.90
System administrator at an East Coast U.S. university	0.90
Systems and security researcher at a Midwest U.S. university	0.91
Ph.D. advisor of Schwartz’s Ph.D. advisor (interested in performance and systems)	0.92
Performance and Systems researcher at a West Coast U.S. university	0.92
System administrator at an East Coast U.S. university	0.92
Head system architect, government research laboratory	0.92

Table 5: Top of Computed Aggregate Specialization Graph Surrounding Schwartz

Applying the algorithm to each of approximately 40 other people whose interests we knew yielded similarly encouraging results. For example, computing the ASG for a researcher at a university at which we did not collect mail logs produced a list containing many people involved with the Computer Professionals for Social Responsibility. This fact corresponds well with one of that researcher’s interests. Running the algorithm on another person at a university where we did not collect data yielded a list of several members of the Internet Architecture Board, of which the person was a member. These examples illustrate the ability of the algorithm to correctly characterize the aggregate interests of a much larger population than just the sites where data was collected. This ability is a practical consequence of the small world phenomenon.

5.2. Magnitude of Interest Distances

The interest distances shown in Table 5 are fairly large. To explore this phenomenon, we computed ASGs surrounding each of 500 randomly sampled nodes in the core inter-domain graph. Figure 4 shows the distribution of the interest distance over the sampled nodes, for nodes other than the distinguished nodes (since distinguished nodes by definition have distance 0 from themselves). The spikes in this graph arose because the core inter-domain graph had many low degree nodes, amplifying distance values that were the ratio of two small numbers ($\frac{1}{2}$, $\frac{2}{3}$, etc.) As can be seen, nodes in an ASG tended to be quite distant from the distinguished node. This observation underscores the complexity of the graph, as it indicates that each user tended to communicate with other users who

communicated with very different sets of users.

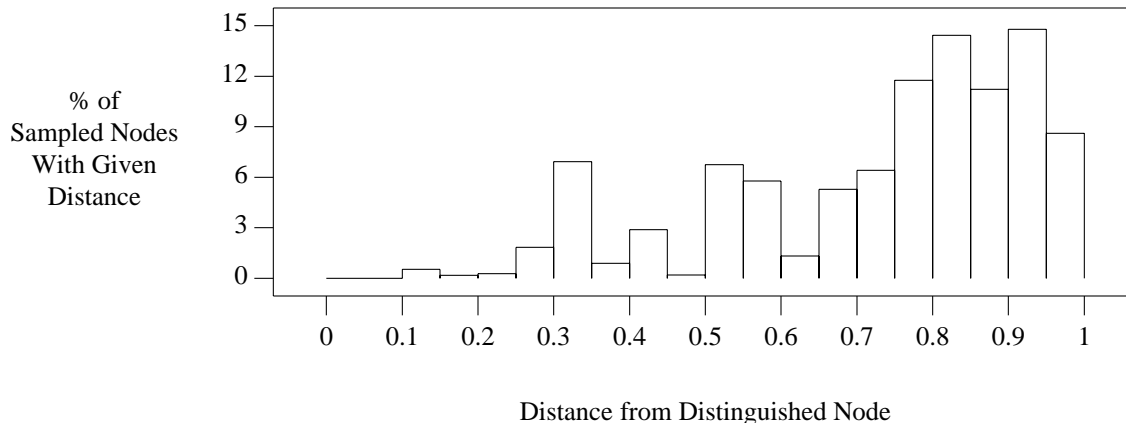


Figure 4: Average Distance Over All Sampled Aggregate Specialization Graphs (histogram)

Extending this observation, we compared the ASG list files for each of the sampled nodes, and found each set to be unique if interest distances were included. When just the names of nodes in each ASG were compared, 205 (0.16%) of the $\binom{500}{2}$ possible pairs were not unique. These measurements indicate that an ASG is like an "interest signature" for a person. This fact is useful in the next algorithm we present.

5.3. Specialization Subgraph Derivation Algorithm

An ASG does not directly indicate how people are related, since it concerns several interests. Yet, by starting with a small number of people known to share a particular interest and constructing a list of individuals high in these lists, one can derive a set of people who likely share that particular interest. To investigate with this idea, we performed three experiments. First, we tried taking the intersection of the individual ASG lists. We found that the success of this approach was heavily influenced by the choice of starting nodes. Starting with two particular nodes, the intersection set in some cases was nearly as large as the individual ASG lists, while in other cases it was nearly empty. Intersecting more than two ASG lists was even more sensitive to the starting nodes.

The problem with the intersection computation is that it is too stringent a restriction to require a person to be in each isolated ASG list when deriving a particular SSG. Therefore, our second experiment started with the set of ASG lists and summed the quantity $(1 - \text{distance})$ across the lists, producing a number for each node corresponding to its overall distance from the nodes closest to the distinguished nodes. This algorithm produced very promising lists of people related to chosen starting people, concerning shared interests in such closely related areas as distributed computing, networks, and naming. By specifying only a few "seed" users, many other highly relevant people were found. Since the analysis in Section 5.2 showed that people have nearly unique ASGs, combining the ASGs in this way from only a few people known to share a common interest is very likely to isolate other people who share that interest.

Our final experiment was an attempt to apply the above SSG derivation algorithm to determine the most pronounced interest for a particular person. Starting from one distinguished person, we generated a list of the closest 100 people to that person. Next, we computed the interest distance lists for each of these 100 people. For each person in these secondary lists, we summed the quantity $(1 - \text{distance})$ across the lists, producing a number for each person corresponding to her/his overall distance from the people closest to the distinguished person. The results of this experiment were quite poor, in that no interest overshadowed all other interests. We conclude that the SSG derivation algorithm is best used to derive an SSG concerning a particular interest, rather than to discover the primary interests of a person.

The full set of graph analysis steps is illustrated in Figure 5. Note that the first seven steps can be computed once, and used for all subsequent interest discovery computations. Only the final three steps need be computed for each discovery computation. These steps completed in approximately 5 seconds per discovery computation on a

Sun 4/110 workstation with 8 megabytes of RAM and a local disk.

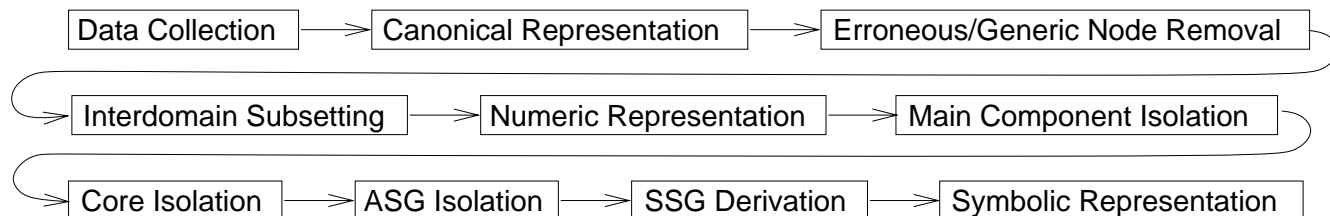


Figure 5: Full Steps Needed For SSG Derivation

5.4. Heuristic Nature of Clustering Algorithms

Clearly, the ASG isolation algorithm and the SSG derivation algorithm do not yield exact results. The experimental ASG lists included some people who, based on our personal knowledge, were incorrectly included in ASGs. Nonetheless, the ASG isolation algorithm can correctly identify a large number of people as ASG members, and the SSG derivation algorithm can derive SSGs based on the ASG isolation algorithm. As such, these algorithms provide a means of extracting organizational hints out of relatively unstructured information. As the networked world becomes increasingly decentralized and heterogeneous, this type of approach can provide a powerful means of tracking changes and distilling order from the potential chaos.

The success of the ASG isolation algorithm is particularly impressive when one considers the modest number of locations at which data was collected. Sampling data from those sites allowed us to deduce hints about shared interest relationships for thousands of people world-wide, spanning many more sites than those where data was collected.

5.5. Robustness to Data Sample

In this section we consider the robustness of the ASG isolation algorithm, as a function of the data sample. We do this by focusing on "star" nodes, or nodes with very high degrees. These nodes represent people who were particularly heavily involved in electronic mail communication, and whose communications were well represented in the sample. The question we wish to answer is how dependent the interest clustering algorithms are on capturing star node communication. If the ASG isolation algorithm can function well even if star nodes are not well represented in the sample, the algorithm is more robust with respect to the data sample.

To measure the importance of star nodes, we derived a subgraph by removing the nodes with degrees of 100 or greater, performed the transformations developed for the original graph (interdomain subsetting, main component location, and core isolation), and applied the ASG isolation algorithm. We found that in a few cases the original nodes to which we applied the ASG isolation algorithm were no longer present in the graph. In the remaining cases, the ASG lists were subsets of the old lists, with closer interest distances. In some cases the order of entries changed. In a few cases the new lists made more intuitive sense. Usually, however, the new lists were somewhat less meaningful, including people who were not as strongly related to the starting person's interests as the people in the corresponding ASG from the original graph. From this experiment we conclude that star nodes are helpful but not critical to the ASG isolation algorithm, and hence that the ASG isolation algorithm is fairly resilient to the data sample.

To better understand the role of star nodes, we observe that there were 564 nodes with degrees of 100 or greater, and that removing these nodes broke the graph into 178 components. Hence, on average each star node "held together" 3.17 components, acting as bridges from smaller subgraphs into the "main" graph. Browsing through the list of star nodes indicated that most fell into one of the following classes, in approximately decreasing order of frequency:

- software maintenance personnel at product development firms
- moderators of large mailing lists
- systems administrators at organizations providing computing support for a number of smaller organizations (e.g., at a university computing center)

- regional network points-of-contact and personnel involved with Internet administration
- department chairpersons and other managers

These observations were drawn from the structure of the mail names, plus our familiarity with many of the star nodes. Note that other than the large mailing list moderators, the categories above are independent of the reasons for forming SSGs. These observations help explain why the star nodes are not critical to the ASG isolation algorithm: while the star nodes perform important functions, their functions are mostly independent of the many SSGs that exist for individual shared interests.

6. Implications and Potential Applications

An important issue raised by the SSG derivation algorithm is the potential privacy threat it poses to electronic mail and other forms of communication. Using the SSG derivation algorithm, it would be possible to deduce shared primary interests simply by monitoring patterns of communication, without access to the text of message traffic. This observation means that, for example, corporations could monitor traffic and generate lists of potential "niche" groups for advertising purposes. Government agencies could also use the technique to generate lists of people potentially associated with particular people under scrutiny. Given the increasing use of telemarketing techniques and the recent increase in government surveillance in cases like "Operation Sun Devil" [Chapman 1990], this is a dangerous possibility. Moreover, it would be difficult to protect against these problems, since doing so would require either encrypting message headers (which could make mail routing difficult) or flooding the network with spurious messages, to make the data "noisier" (which could waste a considerable amount of network bandwidth).

We believe that SSG derivation is inherently neither good nor bad; it is simply powerful, and must be applied ethically. The solution might lie in legal restrictions on the use of clustering algorithms. While we are not legal or policy experts, this approach seems problematic, given the increasing lag between the introduction of a new technology and effective laws governing its appropriate use. The guideline we recommend is that clustering techniques be used only with explicit consent of the users involved. Assuming this guideline, we now consider potential applications that we believe could be carried out without invading privacy.

One interesting possibility would be to use the SSG derivation algorithm to discover users who might be interested in or knowledgeable about a particular topic. Rather than explicitly specifying individual users (the electronic mail paradigm) or news groups on which to post messages (the news paradigm), a user could specify a small set of "seed" people, each of whose interests are believed to be close to a topic of interest. The system could then use the SSG derivation algorithm (applied against the graph monitored at a small number of sites; global data collection is not necessary) to form lists of users whose interests are probably close to those of each of the seed people. The user might then be asked to select among the generated list. This way, shared interests would be specified using the communication graph itself, rather than an artificially imposed classification structure.

As with electronic bulletin boards, SSG disclosure would be voluntary, with people explicitly aware that SSG derivation was being performed. People who believed that allowing access to their SSG information posed too severe a potential privacy threat could choose not to participate. An application that would likely not be privacy-invasive would be providing an implicit index into public technical discussions.

Other forms of resource discovery could also be supported using SSG derivation. One class of potential applications concerns tracking usage patterns in a file system, to provide data about SSGs. The pattern of access of one piece of the file system in relationship to other pieces could provide clues as to relevant information. For example, a file system might contain a series of standards documents, such as the CCITT Blue Book. If a user knows about the X.25 data communication protocol but is interested in other relevant standards, an SSG might point towards related standards, such as the LAPB data link layer standard or the V.32 physical layer standard. SSG derivation would allow these relationships to be built up over time without a requirement that the relationships be identified in advance by an indexer or library administrator. As another example, usage patterns in a local file system could be used to help novice users locate needed information. As a third example, in a system that supports multiple *views* of data, a system could uncover implicit relationships between views, so that users exploring one view could follow links to related views. We are exploring this idea in the context of resource discovery among Internet sites that support the "anonymous" File Transfer Protocol [Schwartz et al. 1991].

Another potential application of the SSG derivation algorithm would be to apply it to data in a corporate environment, to uncover situations in which the corporate organizational hierarchy was inefficient (i.e., where the SSGs did not closely correspond to the established bureaucratic subtrees). The challenge would be to do this

without violating privacy. For example, it might be possible to apply the algorithm on data that has been passed through a trap-door function, such that people's identities are not visible, but working group-level corporate hierarchy interconnections are visible. Clearly, applying this technique should only be done after explaining the procedure to employees, and soliciting their feedback.

7. Related Work

A number of systems have been built to provide directories of computer users [CCITT/ISO 1988, Droms 1990, Harrenstien, Stahl & Feinler 1985, Peterson 1988, Schwartz & Tsigotis 1991] To the best of our knowledge, no one has experimented with trying to support user discovery by observing communication patterns.

Information retrieval techniques have been developed to build dynamic relationships between data, based on automatically extracted keywords about the data [Kahle & Medlar 1991, Salton 1986]. Often, however, doing so presents problems of keywords matching too little or too much information. Moreover, these techniques are typically limited to textual resources. It would be difficult to use them to interrelate users, without inspecting mail and file contents.

The graph analysis algorithms we developed are an example of traffic analysis. Traffic analysis has been used to help cryptanalysts, economists, or military intelligence agencies analyze incomplete data (e.g., to provide clues about the contents of an encrypted data stream) [Callimahos 1989]. In contrast, in the current paper we use traffic analysis to support interpersonal resource discovery.

Some studies have been made of usage patterns in electronic mail systems [Schroeder, Birrell & Needham 1984, Terry 1985]. These studies focus on measurements to uncover how one might improve caching or routing algorithms, while the current paper focuses on characterizing how people cluster by shared interests.

Finally, a number of sociological studies have measured the small world phenomenon in moderate sized communities [Milgram 1977, Travers & Milgram 1969]. The current paper measures this phenomenon in a much larger setting, and applies the idea of a small graph diameter to finding a core subgraph that can be analyzed using traffic analysis to uncover shared interests.

8. Conclusions

In this paper we have focused on the problem of discovering users with particular interests or expertise. Typical directory and electronic mail services address this problem through explicitly registered special interest group lists. Unfortunately, building and maintaining such lists requires a significant amount of effort, and assumes that one knows what lists should be built, and who should be included in each list. We presented an entirely different approach, which uses a set of heuristic algorithms to uncover shared interest relationships between people, based on the history of communications between people.

The algorithms extract implicit organizational structure from the graph. By starting with a distinguished person, the first algorithm we presented can isolate lists of other people, many of whom share some interests with that person. By starting with a small number of people known to share a particular interest and constructing a list of individuals high in these lists, the second algorithm we presented can derive a list of people who share that particular interest. This latter algorithm has powerful potential for supporting resource discovery and various types of collaboration.

The densely interconnected nature of communication graphs (the "small world" phenomenon) allows one to perform shared interest analysis effectively on data collected at even a modest number of locations. We demonstrated the algorithms in this paper by applying them to logs collected from only 15 sites for only two months, yet these data generated a graph containing approximately 50,000 people in 3,700 different sites distributed among 31 different countries world-wide. From this graph we were able to analyze interests of people far beyond the sites where data was collected.

We believe the algorithms we presented are inherently neither good nor bad; they are simply powerful, and must be applied ethically. In particular, one could build a system based on the algorithms we have described, in which users only participate on a voluntary basis. For professional collaborations, people may be interested in having this service with explicit knowledge of its ramifications, just as users of current electronic bulletin boards are explicitly aware that their communications are publically visible. Other forms of resource discovery could also be supported using non-personal resources, for example to interrelate parts of a large set of documents based on the history of access patterns to a file system containing these documents.

Acknowledgements

We would like to express our appreciation to the sites that allowed us to collect data from their systems. We thank Andrzej Ehrenfeucht and Marvin Solomon for suggesting some of the graph computations used in Sections 4 and 5. We thank Rebecca Marvil for helping to implement some of the graph computations. We thank the people who allowed us to use their names in the SSG example discussed in Section 2. Finally, we thank Jonathan Bein, David Goldstein, Goetz Graefe, Carl Malamud, David Wagner, Panagiotis Tsirigotis, and the referees and editors for their comments on this paper.

We used databases from the BITNET Network Information Center, Merit, and the UUCP map to translate host names to geographical coordinates to generate the maps in this paper.

This paper is based on work reported in part in an earlier paper, entitled "A Measurement Study of Organizational Properties in the Global Electronic Mail Community".

This material is based upon work supported in part by NSF cooperative agreement DCR-8420944, and by a grant from AT&T Bell Laboratories.

9. References

[Allman 1985]

E. Allman. *Sendmail - An Internetwork Mail Router*. Comput. Sci. Division, EECS, Univ. of California, Berkeley, June 1985.

[CCITT/ISO 1988]

CCITT/ISO. *The Directory, Part 1: Overview of Concepts, Models and Services*. CCITT/ISO, Gloucester, England, Dec. 1988. CCITT Draft Recommendation X.500/ISO DIS 9594-1.

[Callimahos 1989]

L. D. Callimahos. *Traffic Analysis and the Zendian Problem*. Aegean Park Press, Laguna Hills, CA, 1989.

[Chapman 1990]

G. Chapman. Supporting CPSR's Work in Civil Liberties. Letter from Executive Director of the Computer Professionals for Social Responsibility to CPSR Members, Aug. 1990.

[Droms 1990]

R. E. Droms. Access to Heterogeneous Directory Services. Proc. 9th Joint Conf. of IEEE Computer and Communications Societies (InfoCom), June 1990.

[Harrenstien, Stahl & Feinler 1985]

K. Harrenstien, M. Stahl and E. Feinler. NICName/Whois. Oct. 1985.

[Hoffman 1987]

P. Hoffman. The Man Who Loves Only Numbers. Nov. 1987.

[Kahle & Medlar 1991]

B. Kahle and A. Medlar. *An Information System for Corporate Users: Wide Area Information Servers*. Interop, Inc., Nov. 1991.

[Lottor 1992]

M. Lottor. Internet Growth (1981-1991). Req. For Com. 1296, Network Information Systems Center, SRI Int., Jan. 1992.

[Milgram 1977]

S. Milgram. The Small World Problem. In S. Milgram, editor, *The Individual in a Social World*, pp. 281-295, Addison Wesley, Reading, MA, 1977.

[Mockapetris 1987]

P. Mockapetris. Domain Names - Concepts and Facilities. Req. For Com. 1034, USC Information Sci. Institute, Nov. 1987.

[Peterson 1988]

L. L. Peterson. The Profile Naming Service. *ACM Trans. Comput. Syst.*, 6(4), pp. 341-364, Nov. 1988.

[Postel 1981]

J. Postel. Internet Protocol - DARPA Internet Program Protocol Specification. Req. For Com. 791, USC Information Sci. Institute, Sep. 1981.

[Quarterman 1990]

J. Quarterman. *The Matrix - Computer Networks and Conferencing Systems Worldwide*. Digital Press, 1990.

- [Salton 1986]
G. Salton. Another Look at Automatic Text-Retrieval Systems. *Commun. ACM*, 29(7), pp. 648-656, July 1986.
- [Schroeder, Birrell & Needham 1984]
M. D. Schroeder, A. D. Birrell and R. M. Needham. Experience with Grapevine: The Growth of a Distributed System. *ACM Trans. Comput. Syst.*, 2(1), pp. 3-23, Feb. 1984.
- [Schwartz et al. 1991]
M. F. Schwartz, D. R. Hardy, W. K. Heinzman and G. Hirschowitz. Supporting Resource Discovery Among Public Internet Archives Using a Spectrum of Information Quality. *Proc. 11th IEEE Int. Conf. Distrib. Comput. Syst.*, pp. 82-89, May 1991.
- [Schwartz & Tsirigotis 1991]
M. F. Schwartz and P. G. Tsirigotis. Experience with a Semantically Cognizant Internet White Pages Directory Tool. *J. Internetworking: Research and Experience*, 2(1), pp. 23-50, Mar. 1991.
- [Schwartz 1992]
M. F. Schwartz. Internet Resource Discovery at the University of Colorado. To appear, *IEEE Computer Magazine*, Revised Oct. 1992.
- [Sedgewick 1988]
R. Sedgewick. *Algorithms*. Addison Wesley, Reading, MA, 1988. Second Edition.
- [Spafford 1989]
E. H. Spafford. The Internet Worm: Crisis and Aftermath. *Commun. ACM*, 32(6), pp. 678-687, June 1989.
- [Terry 1985] D. B. Terry. Distributed Name Servers: Naming and Caching in Large Distributed Computing Environments. Ph.D. Diss., Tech. Rep. UCB/CSD 85/228, Comput. Sci. Division, EECS, Univ. of California, Berkeley, 1985.
- [Travers & Milgram 1969]
J. Travers and S. Milgram. An Experimental Study of the Small World Problem. *Sociometry*, 32(4), pp. 425-443, 1969.