A Measurement Study of Organizational Properties In the Global Electronic Mail Community *

Michael F. Schwartz

CU-CS-482-90

University of Colorado at Boulder
DEPARTMENT OF COMPUTER SCIENCE

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

ANY OPINIONS, FINDINGS, AND CONCLUSIONS OR RECOMMENDATIONS EXPRESSED IN THIS PUBLICATION ARE THOSE OF THE AUTHOR AND DO NOT NECESSARILY REFLECT THE VIEWS OF THE NATIONAL SCIENCE FOUNDATION

A Measurement Study of Organizational Properties in the Global Electronic Mail Community[†]

Michael F. Schwartz

David C. M. Wood

CU-CS-482-90 August 1990

Department of Computer Science
Campus Box 430
University of Colorado
Boulder, Colorado 80309-0430
(303) 492-7514

Abstract

Computer systems intended for use in large scale environments are typically organized according to rigid hierarchical structures. For example, traditional file and directory services rely on hierarchical organization to enhance scalability. Motivated by hierarchy's poor support for navigating among large, highly diverse collections of resources (the *resource discovery* problem), we have become interested in organizational structures that arise naturally when people collaborate. In this paper we explore the graph structure resulting from global electronic mail communication. We characterize the structure through analysis of data collected about international electronic mail communication patterns among approximately 50,000 people in 3,700 different administrative domains. We define an *Interest Specialization Graph* structure that provides the scalability of a hierarchy without its organizational inflexibility. We believe that systems organized with this graph structure offer promise of better supporting the organizational needs of a large environment characterized by widespread interorganizational collaboration.

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

1. Introduction

Ongoing increases in wide-area network connectivity enable an interesting set of distributed applications, whereby individuals may more effectively participate in *distributed collaboration*, or the accomplishment of tasks through sharing of resources among many interrelated individuals across administrative boundaries. While intra-administrative domain collaboration is also important, we focus on inter-administrative domain collaboration here, because of the new possibilities that exist to support such collaboration, given recent advances in internetworking.

The primary means of distributed collaboration currently is electronic mail. Yet, the possibility exists for a much more significant degree of sharing. Electronic mail is primarily used in a point-to-point fashion, supporting the interchange of messages between pairs of individuals. We are interested in a more concurrent, symmetrical style of collaboration [Schwartz & Demeure 1989] involving many participants. For example, a more powerful sharing mechanism could be modeled on the types of interactions that take place at conferences and other meetings, where people discuss issues collectively with people they had not previously met, but who, by their presence, are known to have closely related interests.

Our interest in the general area of distributed collaboration concerns the problem of organizing and searching large, shared resource spaces across organizational boundaries. In this connection, the Networked Resource Discovery Project at the University of Colorado, Boulder, is investigating means by which users can discover the existence of a variety of resources in an internet[2] environment, such as retail products, network services, and people in various capacities [Schwartz 1989]. A key problem is organizing the resource space flexibly. While approaches based on a hierarchical organization (such as the CCITT X.500 directory service [CCITT 1988] and Lampson's global name service [Lampson 1986]) have good scalability properties, they are ultimately of limited use for resource discovery. As a hierarchically organized service registers an increasingly wide variety of resources, searching for resources becomes difficult, because the organization becomes convoluted and requires users to understand how its components are arranged. We are exploring several more flexible organizational approaches.

In the current paper we look beyond the resource discovery problem at the more general issue of supporting distributed collaboration. By studying how people collaborate in the global electronic mail community, we illuminate ways in which current computer system organizations are inadequate, and suggest approaches that can be built into applications intended for supporting collaboration and sharing in widely distributed environments.

1.1. Interest Specialization Graph Structure

Our interest in large scale organizational structures was partially motivated by observations about the organization of human social networks. Such networks represent a non-hierarchical structure that scales well. Rather than forming contacts with each other based on a hierarchy, people often establish more direct "networks", by contacting knowledgeable intermediaries who can quickly refer them to other relevant people, cutting across bureaucratic boundaries. For example, by contacting a computer science professor or network manager, someone interested in high-speed networking technology can quickly meet other people who share this interest. These people can, in turn, introduce the person to others who perhaps more closely share his/her particular interests. At the same time, the newcomer can be instrumental in pointing out individuals who share other interests with the people he/she meets.

An interesting characteristic of such networks is what has been called the "small world" phenomenon [Travers & Milgram 1969]. Consider a graph where nodes are people and edges represent one person's knowing another. It has been observed that the diameter of such a graph (i.e., the number of edges in the longest path between any two nodes in a minimum spanning tree of the graph) is surprisingly small, even in an enormous setting. As a small example, there is a mathematical game based on the co-author relationship with the prolific

[2] Throughout this paper we use "internet" to refer to general internetworks. We use "Internet" to refer specifically to the growing collection of government sponsored networks (including NSFNet, ARPANET, and CSNet) and regional networks (such as Westnet and NYSERNet) that interconnect academic, industrial, and government institutions, primarily in the United States.

mathematician Paul Erdos. In this game, an individual's *Erdos number* is defined to be one if that person has written a paper with Erdos, two if they have written a paper with someone who has written a paper with Erdos, etc. Based on this definition, the highest Erdos number known to be possessed by a person is seven [Hoffman 1987].

Lore has it that the diameter of all people in the world is quite small, perhaps 6 or 7. There have been a number of small-scale sociological studies of this phenomenon [Boissevain 1974, Travers & Milgram 1969], although no one has measured the diameter of a large society. Among other results, the current study provides an indication of this phenomenon over a much larger sample size than previous studies. This phenomenon indicates the existence of a graph structure that is densely interconnected, and that has many different short paths between any two nodes in the graph.

From the perspective of distributed collaboration, an even more interesting characteristic of human social networks is their flexible organizational structure. Rather than forcing all internode relationships to conform to a hierarchy, the graph structure allows individuals to be related to one another through multiple groupings that we call *Specialization Subgraphs (SSGs)*. An SSG is a subset of nodes that share common attributes, and that are closely interconnected (i.e., the subgraph has a small diameter). As an example, in a graph of relationships among people, one SSG could connect individuals based on a shared interest in a particular computer science speciality, a second SSG could connect individuals based on shared responsibilities at a place of employment, and a third SSG could group individuals based on shared cultural/recreational interests. Any individual can belong to many different SSGs, and can search for information about a particular topic by consulting the appropriate SSG. We refer to the overall graph structure represented by these relationships as an *Interest Specialization Graph (ISG)*.

As a concrete example of the ISG structure and its potential efficiency, one of the current paper's authors (Schwartz) recently became interested in finding measurements of the number of naming domains in the Internet [Mockapetris 1987], to help estimate the scope of an Internet directory facility that was built in conjunction with another part of the Networked Resource Discovery Project [Schwartz & Tsirigotis 1990]. One way to locate such measurements would have been to broadcast an announcement requesting information on an appropriate electronic bulletin board. However, doing so reaches a very broad group of people, reducing the likelihood of useful responses, because of the necessarily high overall volume of information flow and the range of knowledge of the potential respondents. Instead, Schwartz chose to send electronic mail messages to a small number of people in what amounted to a relevant SSG. For the sake of concreteness we have included the names and affiliations of the people contacted in Table 1. The reasons for choosing these people are given in the "Relevant Interest" column in the table. The names of a few other people arose during this process, but those people could not be reached to ask for permission to use their names in this description, so they have been omitted from the table.

Individual	Affiliation	Relevant Interest
Phil Karn	Bell Communications Research	Networking Researcher
Sol Lederman	DDN Network Information Center	Network Information Center Technical Staff
Mark Lottor	DDN Network Information Center	Author of "Zone" Program
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 Researcher
Marshall Rose	Performance Systems International	Manager, NYSERNET/PSI White Pages Pilot Project
Karen Sollins	MIT Laboratory for Computer Science	Naming Researcher

Table 1: People Contacted in Example Resource Discovery Process

In the resource discovery process represented by this figure, Schwartz initially sent messages to Karn, Lederman, Mockapetris, Peterson, Rose, and Sollins. In response, Rose, Karn, and Mockapetris all told Schwartz that Lotter had written a program called "Zone" for this purpose. (Schwartz did not know Lottor at this time.) Rose also referred Schwartz to Mockapetris. Lederman suggested that Schwartz send a message to the USENET "comp.protocols.tcp-ip" electronic bulletin board, asking for pointers to such programs. This bulletin board is read by individuals interested in TCP/IP and related network protocols. Sollins referred Schwartz to Mockapetris and to Patton. Patton referred Schwartz to Lottor. Peterson suggested that Schwartz send a message to the "namedroppers" mailing list, asking for pointers to such programs. This mailing list is distributed to individuals interested in naming systems. It represents a SSG intermediate in size between the group Schwartz chose and the comp.protocols.tcp-ip bulletin board. Schwartz then contacted Lottor, who provided the desired measurements.

The SSG revealed from this process is illustrated in Figure 1. In this figure, a directed edge from node x to node y indicates that "x knows about y" could be determined from this set of conversations. From previous knowledge of these people, we are aware of other arcs that could be placed in this graph. These arcs are not included here because we wish to show only those edges that can be deduced directly from this single resource discovery process. Arcs indicating membership in other SSGs (e.g., concerning other technical specialities or hobby interests) have also been omitted.

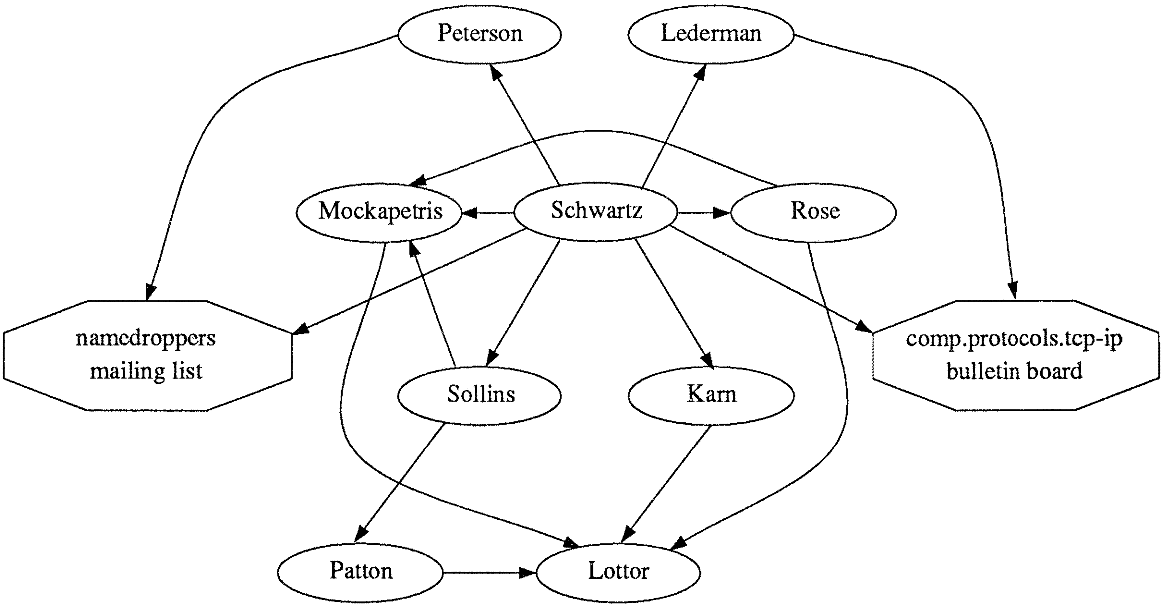

Figure 1: Specialization Subgraph Revealed from Example in Table 1

This example illustrates several properties of ISGs. First, each person contacted shared relevant interests with the others. Second, these interests were more specialized than the range of interests represented by the namedroppers mailing list and the comp.protocols.tcp-ip bulletin board. Third, each person contacted had other interests and responsibilities than those related to the purpose of the query, and hence belonged to other SSGs. This flexibility stands in contrast to the relationships that could be represented in a simple hierarchy, since in that case individuals could only belong to a single organizational grouping. Finally, while a hierarchy typically changes very little structurally over its life, the edges in an ISG are free to change as peoples' interests evolve.

Typically, computer systems do not support this graph structure. Instead, computer system organizations typically fall somewhere in a spectrum between "flat" and hierarchical. "Flat" name spaces (i.e., unordered sets) impose no constraints about how nodes can be related. Such a structure provides flexible organization but poor scalability, since search operations cannot easily be constrained to subsets of such graphs. At the opposite organizational extreme, a hierarchy achieves scalability at the cost of severely limiting organizational flexibility. The ISG structure preserves the scalability of a hierarchy by grouping entities into local *neighborhoods* that limit the scope of operations, without constraining the nesting relationships of these neighborhoods. Moreover, nodes can belong to a variety of different neighborhoods, and unrelated neighborhoods can overlap in arbitrary ways. What is lost is the simple structural characterization that makes it easy to reason about hierarchy. Nonetheless, we believe that application-level support for ISGs can provide much more useful systems for people involved with distributed collaboration and, more generally, for sharing resources in a large scale distributed environment.

1.2. Focus of this Paper

The basic premise of this paper is that sophisticated resource sharing and distributed collaboration (including the resource discovery problem) require organizational support tailored to the ways people naturally communicate, rather than to the simple hierarchical structures typically utilized by large scale computer systems. Our perspective is that large scale distribution is essentially a societal phenomenon, and as such, supporting the necessary organizational structure will require understanding the nature of the ISG structure.

In this paper we characterize the structure of ISGs, focusing on properties that could support flexible and scalable distributed collaboration. Our characterization is necessarily statistical in nature, because it is based on measurements of a large, real-world graph, rather than on an algorithmically generated graph. The graph was derived from information collected about electronic mail usage at a number of universities and companies involved with a variety of research, education, and product development projects. In particular, we collected "From:" and "To:" lines from electronic mail log files on a temporary basis at 15 representative administrative domains around the U.S. and Western Europe. The graph generated from this data contains approximately 50,000 users in 3,700 different administrative domains, with 183,000 edges (i.e., indications of pairs of users communicating via electronic mail).

Throughout this paper we use the term "administrative domain" in reference to the boundaries of free sharing that one typically associates with large scale distributed environments. Such boundaries often exist at the granularity of academic departments, research laboratories, and companies. For machines named according to the Internet Domain Naming System, an administrative domain corresponds to a collection of machines whose names appear as leaves under one node, such as "colorado.edu".

Our results fall into two main parts. First, analysis of the electronic mail communication graph indicates a structure that is naturally redundant (and hence resilient to various types of failures), as well as densely interconnected (and hence an effective conduit for the dissemination of information). This structure is interesting because it provides often-sought properties with considerably more flexibility and less complexity than typical computer systems do. Second, our analysis indicates that people naturally group into a large number of different shared interest groups, with a relatively small number of people relevant to any particular topic. Given a means to chose among shared interests without imposing an artificial organizational structure, the resilience and efficiency of the ISG structure offer the possibility for very effective distributed collaboration. In this connection, we have developed an algorithm to cluster people according to shared interests, based on analysis of the communication graph. We later discuss potential applications of this algorithm, as well as possibilities of privacy invasion from its abuse.

The remainder of this paper is organized as follows. In Section 2 we review related work. In Section 3 we discuss the methods used in collecting the data. In Section 4 we present geographical maps of the data. In Section 5 we present and interpret measurements of the data. In Section 6 we consider possible applications of the analysis done in this study. Finally, in Section 7 we offer our conclusions.

2. Related Work

There have been a number of efforts to support distributed collaboration beyond the basic support of electronic mail. Electronic "bulletin board" services such as USENET [Quarterman & Hoskins 1986] support a richer sharing structure than electronic mail, since individuals may participate in discussions of common interest with thousands of other individuals around the world. Yet, their restrictive organizational structure (a small number of relatively statically defined interest groups) and means of information distribution (unsequenced, full scale broadcast) do not readily facilitate high quality collaboration. Other related work includes hypertext systems [Conklin 1987], community information services [Schatz & Caplinger 1989], whiteboards [Donahue & Widom 1986], digital library systems [Arms 1989, Kahn & Cerf 1988], and the Wulf's "Collaboratory" vision

[Wulf 1989]. Each of these projects focuses on providing tools for resource and information sharing. In contrast, the current paper considers the organizational properties that arise naturally during large scale collaboration across administrative boundaries. Estrin studied the extent to which interorganizational collaboration takes place, as well as some of the issues that arise due to interorganizational communication [Estrin 1985]. In a sense, the current paper extends that work, providing considerably more detailed analysis about the nature of large scale distributed collaboration.

3. Experimental Methods

In this section we discuss the experimental methods employed in our study. We begin by discussing our choice of measurement data, and how these data were collected. We next consider privacy and security issues raised by the study. We then discuss the range of administrative domains that participated in the study. Finally, we discuss the steps used to transform the raw data that were collected into data that could be used in the measurement process.

3.1. Data Collection Choice and Methodology

Our choice of measurement data was governed by the desire to measure a real, working environment that exhibited the organizational characteristics of ISGs. We considered measuring usage patterns in USENET, the Domain Naming System, and electronic mail traffic. We chose to measure electronic mail traffic because mail directly represents interactions between individuals, according to common interests. In contrast, USENET only allows a very rough grained organization, separating discussions into only approximately 600 "news groups" representing shared interests. The Domain Naming System activity only represents host-to-host communication patterns, rather than interest-specific communication.

To collect the electronic mail data, we constructed a pair of program scripts, the first of which 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 automatically ran each day for the duration of the study using the UNIX[3] "at" facility to repeatedly schedule itself at a preset time. This script collected "From:" and "To:" lines recorded in the UNIX "syslog" log files for each message transferred by "sendmail", the Berkeley UNIX mail transfer agent [Allman 1985], filtering out irrelevant information to reduce Internet bandwidth needs. The script enforced an upper bound on the number of bytes that would be sent, because some sites had very large log files. The resulting data were then broken into parts small enough to fit through Internet mail, marked so that they could be resequenced upon receipt. The receiving script made some basic data validity checks, and saved the data in appropriately named files, based on the name of the data collection site the date, and the message sequence number.

Before starting the data collection period, we looked at several hosts at the University of Colorado, to understand the content of the data, and to gauge the network and disk space requirements of collecting the main body of data. We estimated the disk space requirements, network load, and computational requirements on both the sending and receiving hosts. We presumed that as the data collection progressed, the number of new edges noted would decrease as a percentage of all the edges noted each day. We estimated this decrease to be in the same proportion as the number of duplicates found in one day's sample. Using this information and estimating 1600 messages per day from 20 sites, we projected that the data would contain about 20,000 nodes and 113,000 edges. This estimate was low by about a factor of approximately 2.5 in the number of nodes, and 1.5 in edges.

In total, we collected data for an average of 1.5 months per log host, between December 1988 and January 1989. While this duration meant that we missed data for infrequently exchanged messages, it would have been infeasible to collect the data for significantly longer (e.g., a year), in terms of the continued effort and good will of the participating sites, the disk space to store the added data, and the CPU time to analyze the added data. We also note that after only a few days of data collection, the number of new unique edges collected per day

[3] UNIX is a trademark of AT&T Bell Laboratories.

decreased dramatically, to approximately 10% of the initial days' collections.

3.2. Privacy and Security Issues

Privacy and security were important considerations in obtaining study participants. In soliciting participation, we promised prospective sites that we would not reveal any identifying details of the data we collected. Even so, several administrative domains declined participation, citing privacy concerns. Simply monitoring at "To/From" pairs might provide sensitive information beyond the bounds of our study, particularly in cases where the mail was exchanged between corporations. (The interested reader is referred to the traffic analysis literature, e.g., [Callimahos 1989].)

As one possible answer to these privacy concerns, we considered using a mapping of electronic mail addresses to nameless unique identifiers, based on a "trap-door" function (i.e., one in which the results of the translation do not yield information about the inverse mapping). We chose not to use such a scheme for several reasons. First, it would have added computational expense, and we wished to minimize such expenses at the participating sites. Second, for efficiency's sake this mapping would have required the remote sites to run a compiled program, rather than an interpreted script. We wished to avoid such added complexity, because it would present more potential security risks in the eyes of the systems administrators at the participating sites. Third, the transformation would need to accommodate heterogeneity in electronic mail naming. It was not until we had received the data that we knew the full extent of this heterogeneity. Finally, having access to the actual electronic mail names was helpful for interpreting the data, as will be discussed in Section 5.3.

Our ability to obtain study participants was complicated by the Internet worm of November 1988, which invaded the Internet two weeks before we had planned to start our data collection [Spafford 1988]. This incident raised the security consciousness of most system administrators, and also increased their workload at the time. Still, no site withdrew its offer to participate after the worm, and we simply delayed the start of the study period to December 1, 1988.

3.3. Study Site Characteristics

We asked 62 different administrative domains to participate. Among these, fifteen sites containing 22 machines that logged mail traffic (hereafter referred to as "log hosts") did so. Some sites transmitted data from multiple log hosts corresponding to separately administered computing facilities. For example, at the University of Colorado, one log host collected information for departments whose research computing was managed by a contracted computing management service, and a second log host collected information for instructional computing machines administered by the university academic computing facility. Of the 47 sites that did not participate,

- 20 sites gave no conclusive response, after a few gentle "prods"
- 13 sites had administrators who said they were too busy
- 8 sites expressed security concerns
- 4 sites had data that would not be very useful (e.g., too few hosts)
- 2 sites expressed privacy concerns.

The participating administrative domains had the following 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 had 50-200 people
- 7 had 200-1,000 people
- 4 had 1,000+ people

Clearly, collecting data from only 22 log hosts at 15 administrative domains does not constitute a large percentage of the networked population. Yet many organizations (such as the engineering college at the University of Colorado) pass or log much of their mail through a small number of central gateway machines, particularly for mail going outside the organization. Moreover, because the participating administrative domains were geographically and organizationally well distributed, we managed to capture a snapshot of communication patterns at quite a large proportion of the global electronic mail community. Indeed, the overall graph generated by the data collection contained 50,834 users on 17,312 hosts in 3,739 administrative domains.

3.4. Raw Data Transformation Steps

Several operations were needed to transform the collected data into a form that could be analyzed. The implementation of these steps was rather involved, because the data contained a variety of formats and flaws.

The first step involved extracting valid edges from the sendmail log entries. There were two types of complications. First, a wide variety of host naming conventions exists. It was necessary to transform names to a single canonical format to reduce the number of aliases treated as separate nodes. Part of the problem was that a number of nested components may be included in electronic mail names (such as "user%host@host1.domain"). Such components are interpreted in a sequence stages by a number of mail message transfer agents. While we attempted to parse as many such formats as possible, in some cases it was not possible to guarantee that non-distinguishable aliases were being used because of the decentralized nature of interpreting mail names. Another problem was that some sendmail processes generate log entries using relative host names such as "schwartz@latour" or "latour!schwartz", rather than globally meaningful domain-style names such as "schwartz@latour.colorado.edu". It was necessary to transform host names into globally meaningful formats where possible, based on knowledge of where each log file was collected. Another problem was that some names were invalid, and had to be discarded. One particularly bad example had the form "<user%host1@host2.inst.edu>edu>host3%host@domain".

The second type of complication was the fact that the "From:" and "To:" lines of each mail message are logged in separate records that are interleaved with other entries when multiple sendmail processes attempt to deliver mail concurrently. Therefore, it was necessary to match the two halves of each edge. Moreover, mail delivery attempts often fail several times before either successfully completing or being abandoned, with intermediary attempts recorded in the log. It was thus necessary to determine which transmissions completed successfully before deciding that an edge belonged in the graph. Since mail transmission attempts sometimes span several days, we needed to correlate the data across multiple log file entries for each host. Note that it is possible that a correctly addressed message that could not be delivered because of network problems would not be included in the graph because of this restriction. Because it is not possible to distinguish this case from an erroneously addressed recipient, we discarded such edges.

After generating the edge list, the next step was to remove edges that represented irrelevant or spurious shared interest relationships. For example, edges corresponding to mail between any individual and the "operator" on any host were discarded, since the operator login is used as a generic computing environment maintenance function shared by multiple individuals, as opposed to the shared interest relationships we sought to measure. Similarly, edges corresponding to mail sent to the process collecting the data for the current study were discarded. We also discarded edges with nodes whose address started with "news@", since such addresses represent mailing to one of many possible particular news groups without the ability for the study monitoring software to distinguish between news groups representing different interests (that information is contained inside the message body). We did not reject 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 such as these were checked.

Eliminating data may have had the effect of eliminating some relevant traffic since, for example, some system administrators do all of their work as "root". However, discarding edges simply reduces the size of what was already a statistically sampled subgraph of the global electronic mail communication graph.

In total, we collected logs concerning 1,234,862 electronic mail messages, and removed 203,636 (16.49%) of them because of problems as described above. An interesting peripheral fact is that 130,275 (63.97%) of the bad messages were not unique, indicating that a large amount of global processing effort is being expended trying multiple times to deliver erroneously addressed messages.

Once these transformations had been done, we generated a numeric representation of the graph, to allow efficient graph computations. For example, the graph

user1@host1	user2@host2
user2@host2	user3@host3

would be transformed to the graph

```
1  2
2  3,
```

plus a table listing the transformations, to allow the results of graph computations to be mapped back into symbolic form.

4. Geographical Maps

While geographic maps of the collected data are not necessary for analyzing the organizational properties of distributed collaboration, such maps are interesting to view. Figure 2 shows an international map of the edges in the measured traffic that could be plotted. We used a number of databases that translate from host names to geographical coordinates to generate this map. The BITNET Network Information Center provided a database listing the coordinates for many of the BITNET sites. The UUCP map (available in the news group "comp.mail.maps") provided similar information for UUCP sites. For sites directly connected on the IP portion of the Internet [Postel 1981], Merit (the corporation administering NSFNet) provided a mapping from IP network number to geographical coordinates. Because many sites were not BITNET, UUCP, or IP sites, and because some of the databases were incomplete, we could determine the geographical coordinates for only 50% of the data points.

As can be seen, the study captured traffic among Australia, Canada, China, Western Europe, Israel, Japan, Mexico, South America, and the U.S. The U.S. appears as a focal point in electronic mail communication because 13 of the 15 data collection sites were located in the U.S., and because the U.S. currently has more networked sites than other countries in the world. We attempted to plot a "zoomed in" map focusing on the U.S., but the result was covered with so many edges that it conveyed very little information. However, Figure 3 shows the U.S. map with no edges, more clearly indicating where the sites were that are represented in the data. In this map it is interesting to note the nearly vertical line of points dividing the plains from the eastern portion of the country. Notice also that nearly every state is represented in this map. Also, this map looks similar to (although more detailed than) the map of networks joined by NSFNet. This observation underscores the importance of a national network infrastructure for supporting distributed collaboration. To some extent these maps show areas of concentration of our data collection, yet they also show that our data spans a much broader range than just the data collection sites. For instance, we did not collect any data in Boston, yet Boston is heavily represented in Figure 3.

5. Measurements

In this section we present measurements of the collected electronic mail data. We begin by characterizing the graph at a macroscopic level of detail, indicating the size of the graph, the graph diameter, average path length, and a division between a tightly connected core subgraph and a surrounding loose, web-like structure. We then consider a microscopic level of detail, beginning with the distribution of node outdegrees, and progressing to consider the types of nodes that belong to different parts of this distribution. We then introduce a graph measure of intermediate scope, to isolate nodes that share closely related interests, and analyze the ways in which people fit into groupings of other individuals with related interests. This analysis characterizes how people collaborate in a large scale distributed environment.

Throughout the remainder of this section we consider several levels of abstraction of the electronic mail graph. At the finest level of detail, nodes were individual users, and edges indicated communication between particular users. We call this the "User" level graph. At the next coarser level of detail, nodes were hosts, and edges indicated communications between particular users on those hosts. We call this the "Host" level graph. At a still coarser level of detail, nodes were administrative domains, and edges indicated communications between users at these domains. We call this level the "AdmDom" level. More precisely, these nodes were only *apparent* domains, because we used a simple heuristic to determine the domain associated with a particular host

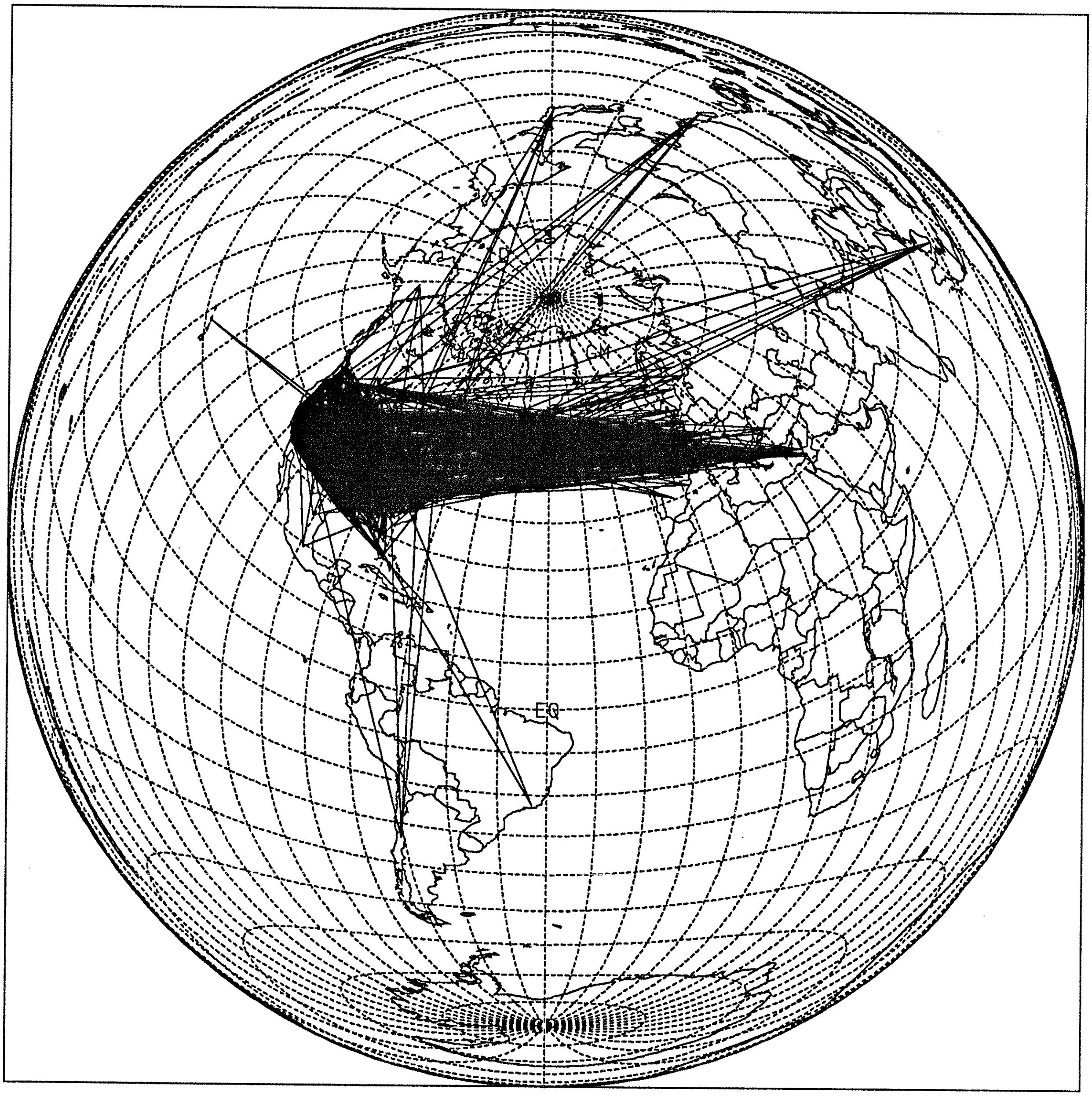

Figure 2: Mappable Subset of Recorded Global Electronic Mail Edges

name. The basic algorithm involved removing the first component of the node name (e.g., yielding "colorado.edu" from the name "latour.colorado.edu"). This algorithm does not always work correctly because some domains provide mail aliasing mechanisms whereby mail can be addressed to "user@domain" rather than "user@host.domain". In such a case, stripping the first component will lose some information. If the domain name is sufficiently deep (such as "cs.washington.edu"), this loss of information will not be critical, since names will still be divided roughly meaningfully. However, for shallow administrative domains like "mit.edu", stripping the first component would provide too course grained a classification to be meaningful. Therefore, the algorithm we used was to strip the first component for names with at least three components. There is no better means for determining the domain associated with a host. Because of the decentralized administration of the

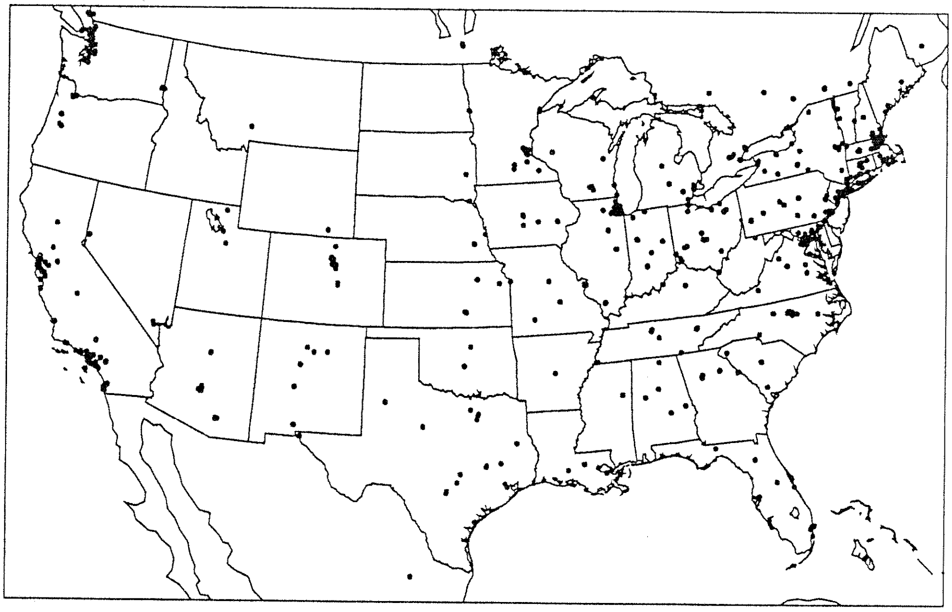

Figure 3: Data Points from U.S. Portion of Figure 2

Domain Naming System, a list of all Internet domains does not exist.[4]

In addition to these levels of detail, we considered a graph derived by removing the host names from the User level graph. We refer to this graph as the "MinusHost" graph. The purpose of generating this graph was to measure the effect on the graph of making the (somewhat optimistic) assumption that user names are unique within an administrative domain. This graph allows us to measure the effects of situations where, for example, user1 sends mail to user2 from host1, and then to user3 from host2. In the basic User level graph, these two edges would be placed in two disjoint subgraph components. By removing the host names, the graph becomes a single component of diameter 2. While the assumption of administrative domain-wide unique user names is not always true, this measurement shows the possible extent to which the graph measurements we made incorrectly caused the graph to look larger and less connected than it really is.

From the MinusHost graph we derived another graph by including only inter-administrative domain edges (e.g., "user1@colorado.edu - user2@mit.edu"). We refer to this graph as the "InterDom" graph. This graph allows us to remove many of the edges that correspond to local administrative communication, rather than true shared interests. The InterDom graph will be considered in more detail in Section 5.3.

Finally, we considered a subgraph of the User level graph where edges were included only if a *pair* of messages, one in each direction, occurred in the data. We refer to this as the "TwoWay" graph level. This graph eliminates edges that might not indicate shared interests between individuals. If a pair of nodes is joined by edges in both directions rather than just one direction, the people represented by those nodes are more likely to know one another. For example, this graph eliminates edges corresponding to messages that are sent to groups of recipients, some of whom are not known by the sender, and do not truly share interests in the subject of the message. While mail addressed to mailing lists often represents shared interest relationships, mailing lists such as "staff" are so general-purpose that the relationship between sender and recipient is questionable. The TwoWay graph provides a baseline comparison for the scope of people who more likely know and share interests

[4] The list generated by the Zone program mentioned in the specialization subgraph example given in Section 1 is incomplete, because some administrative domains do not allow their data to be retrieved across administrative domain boundaries.

with each other.

5.1. Macroscopic Graph Measurements

Table 2 reports the basic dimensions of the electronic mail graphs, at each level of abstraction. The administrative domain count is not completely accurate, because of the simple heuristic we used to determine the administrative domain associated with a particular host name. The interconnection density metric is the ratio of the number of edges in the graph to the number of edges in a fully connected graph of N nodes. Clearly, since the graphs are so large, they are bound to be very sparsely interconnected. Therefore, a more meaningful interconnection density metric is the ratio of E/N, which is equivalent to the average outdegree. We will return to these metrics later.

The node and edge counts in Table 2 indicate that there are 2.94 users per host and 4.63 hosts per administrative domain. Clearly, these figures are quite small, since they only represent statistical samples of the full graph. Moreover, the latter figure is somewhat misleading, because of the simple heuristic we used to determine the administrative domain associated with a particular host name. The edge count from User to Host level only reduced by a factor of 1.42 (which is less than the node reduction factor), indicating that even when users shared a host they often communicated with other users on different hosts. This point reflects the workstation oriented nature of the computing environment. In contrast, the edge count from host to AdmDom level reduced by a factor of 11.69, which is larger than the node reduction factor. This observation reflects the fact that a large amount of communication takes place within an administrative domain. Finally, note that the MinusHost level graph had approximately 81% as many nodes as the User level graph, indicating that as many as 19% of the nodes considered unique in the User level graph might be equivalent to other nodes in that graph level. This fact is the reason why we derived the InterDom graph level from the MinusHost graph level, for use in measuring shared interests between users in Section 5.3.

Level	Example	Nodes (N)	Edges (E)	Interconnection Density: 2E/N(N-1)	E/N
User	user1@mach1.dept1.dom1.edu- user2@mach2.dept2.dom1.edu	50,834	183,833	1.42×10^{-4}	3.62
MinusHost	user1@dept1.dom1.edu- user2@dept2.dom1.edu	41,104	155,788	1.88×10^{-4}	3.79
InterDom	user1@dept1.dom1.edu- user2@dept2.dom2.edu	23,659	43,620	1.56×10^{-4}	1.84
Host	mach1.dept1.dom1.edu- mach2.dept2.dom1.edu	17,312	129,271	8.63×10^{-4}	7.47
AdmDom	dept1.dom1.edu- dept2.dom1.edu	3,739	11,058	1.58×10^{-3}	2.96
TwoWay	user1@mach1.dept1.dom1.edu- user2@mach2.dept2.dom1.edu, user2@mach2.dept2.dom1.edu- user1@mach1.dept1.dom1.edu	9,026	18,768	4.52×10^{-4}	2.08

Table 2: Graph Level Dimensions

The next step we took in characterizing the data was to count disjoint components. It turned out that each graph level contained a surprisingly large number of components. Upon further investigation, we determined that at each level there was one very large component (which we refer to as the *main* component) and a large number of very small components. The User level graph measurements are shown in Table 3. Looking at a sampling of the individual nodes comprising the smaller components in the User level graph indicated that the small components were mainly comprised of two classes of users. The first were BITNET users [Quarterman & Hoskins 1986], reflecting the fact that BITNET users did not have as easy access to wide-area network electronic mail service as do other users of the Internet.[5] The other main class of users in the small components were non-computing professionals, whose connections to the "mainstream" electronic mail community appeared to be only through a small number of other non-computing professionals, plus the systems administrators who set up their accounts. For example, this applied in one case to astrophysicists at two universities communicating with one another.

The remaining measurements we made in the current section were done on the main component of the User level graph, since that graph provides the most detailed opportunity for characterizing the relevant structural properties of ISGs. In the following discussion, we refer to the main component of the User level graph as simply "the graph".

Number of Components		Main Component Size		Second Largest Component Size	
All	Single Edge Comps.	Nodes	Edges	Nodes	Edges
2,557	1,664 (65.1%)	43,498 (85.6%)	179,030 (97.4%)	28 (0.1%)	31 (.00017%)

Table 3: User Level Graph Component Sizes

After completing the basic measurements above, we used a breadth first search algorithm to compute the diameter of the graph. The diameter turned out to be 21. While this value is fairly small relative to the graph size, it is larger than we had expected, based on the "small world" lore discussed in Section 1. We hypothesised that the graph contained a sizable subgraph with a much smaller diameter, surrounded by a relatively sparse set of edges caused by the fact that we only collected data from a fairly small subset of log hosts. To test this hypothesis, we wanted to measure the average path length between each pair of nodes. Because performing this computation requires $O(N^2+NE)$ operations for a graph with N nodes and E edges [Sedgewick 1988], we chose instead to compute the average path length from each of a randomly selected subset of nodes to all other nodes in the graph.

To ensure that the sample selection process was statistically valid, we performed the random sampling on a range of subset sizes, from 1 to 200. We found that once more than 10 nodes were sampled, the average path length varied by very little. The results indicated that the average path length between any two nodes was indeed significantly smaller than the graph diameter. While performing these computations we also computed the sampled maximum path length. These measurements are summarized in Table 4. These three values demonstrate that the shortest path between any two users is quite small on average, even though the worst case path is significantly larger. Moreover, on average nodes are significantly closer to even their most distant counterparts than the graph diameter would indicate, hinting that a collection of nodes around the outer part of the graph contribute substantially to diameter without affecting the path lengths between many of the other nodes.

Diameter	Sampled Average Path Length (accuracy)	Sampled Maximum Path Length (accuracy)
21	5.96 (+/- 6.6%)	13.99 (+/- 2.4%)

Table 4: Diameter and Sampled Path Lengths of Main User Level Graph Component

To further characterize the basis for the smaller diameter "core" subgraph, we hypothesized that a substantial proportion of the graph diameter was caused by nodes with very low outdegrees, reasoning that the core subgraph was surrounded by a loose, web-like structure of people who communicated with only a few other individuals (or those for whom our data collection sites only managed to capture a small proportion of mail). To test this hypothesis, we iteratively constructed a sequence of subgraphs of the main graph component, each obtained by removing all of the nodes with degree 1 from the previous subgraph, until no such nodes remained. The

[5] The data for this study were collected before BITNET merged with CSNet.

algorithm reduced the graph as charted in Table 5.

Iteration Number	Nodes (% Original Size)	Edges (% Original Size)
0	43,498 (100%)	179,030 (100%)
1	21,689 (50%)	157,221 (88%)
2	20,463 (47%)	155,995 (87%)
3	20,161 (46%)	155,693 (87%)
4	20,084 (46%)	155,616 (87%)
5	20,058 (46%)	155,590 (87%)
6	20,051 (46%)	155,583 (87%)
7	20,050 (46%)	155,582 (87%)

Table 5: Main User Component Core Subgraph Isolation Results

The number of iterations executed roughly corroborates our hypothesis, hinting that there is 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). This outer structure could likely have contributed to much of the difference between the sampled average path length and the diameter of the graph. Also, the fact that most of the outer structure was removed during the first iteration of the graph pruning algorithm indicates that there is a crisp distinction between the core graph and the outer structure. This observation indicates that the members of the outer structure are peripheral to most of the communication taking place in the core graph (rather than just sparsely sampled). Another indication of this fact is that 46% of the nodes in the graph were responsible for 87% of the communication that took place (see Table 5). Because of this communication focus, the core graph could be analyzed without some of the problems that arise due to uneven coverage, as will be seen in Section 5.3.

Note that we chose not to consider edge weights (i.e., message transmission counts between pairs of nodes) in the core graph isolation algorithm because of the relatively short duration of the data collection period.

Figure 4 shows a map of the U.S. portion of the core graph. Since the edges in this map connect a proportionately smaller number of nodes, the graph appears significantly less crowded than the full U.S. edge map. In fact, this map appears to have "hubs" at many of the sites where we collected data, confirming the appropriateness of this graph reduction transformation.

In summary, the macroscopic graph measurements in this section indicate that the graph derived from the collected electronic mail logs consists of a large number of singleton components, and one very large component. The main component has a fairly small diameter and a small average path length between nodes. This property enhances rapid and reliable dissemination of information.

5.2. Microscopic Graph Measurements

For a microscopic analysis of the graph, we measured node outdegrees. Our initial attempts at plotting this data indicated that the outdegree distribution has the form of an exponential decay over a wide range of X and Y axis values, yet using a log-log scale yielded a scattered looking graph, because it accentuated the fluctuations of the individual outdegree counts. To render a more intuitively appealing plot, we observed that there are actually three regions of interest: a large number of nodes have a low outdegree; a moderate number of nodes have a high outdegree; and a large number of nodes lie in the middle region. Figure 5 plots the distribution broken into these regions. Plotting the data in this fashion is somewhat like using an logarithmic scale, since the scale changes dramatically between subplots, yet this plotting technique removed much of the scatter seen in the exponentially plotted graph. For this figure we chose regions based roughly on order-of-magnitude decreases in the percentage of nodes at each outdegree. It turned out that these break points occurred near outdegrees of 10 and 100. We chose 10 and 100 as the break points (rather than the true breakpoint values) to yield more intuitively appealing sub-plots.

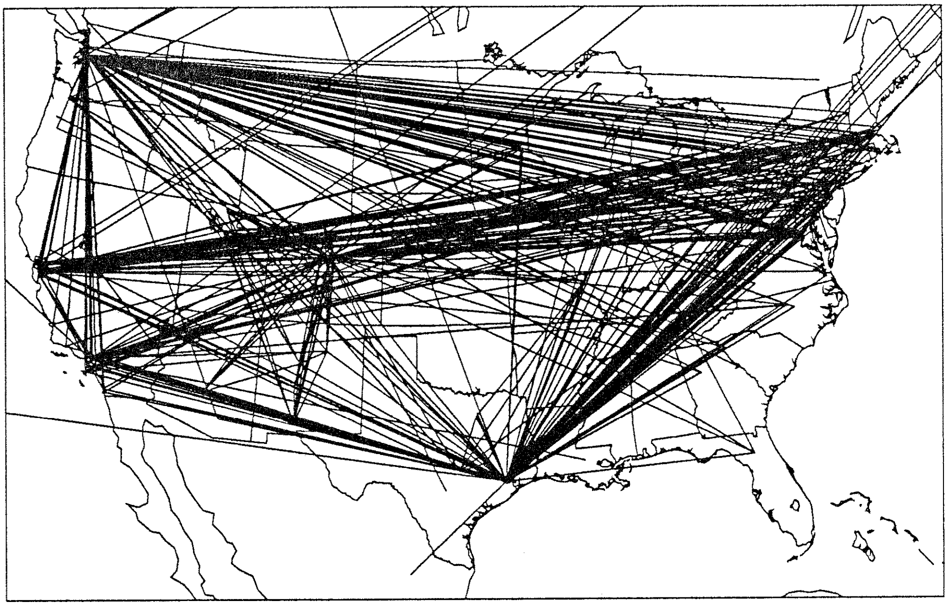

Figure 4: Core Graph of U.S. Portion of Figure 2

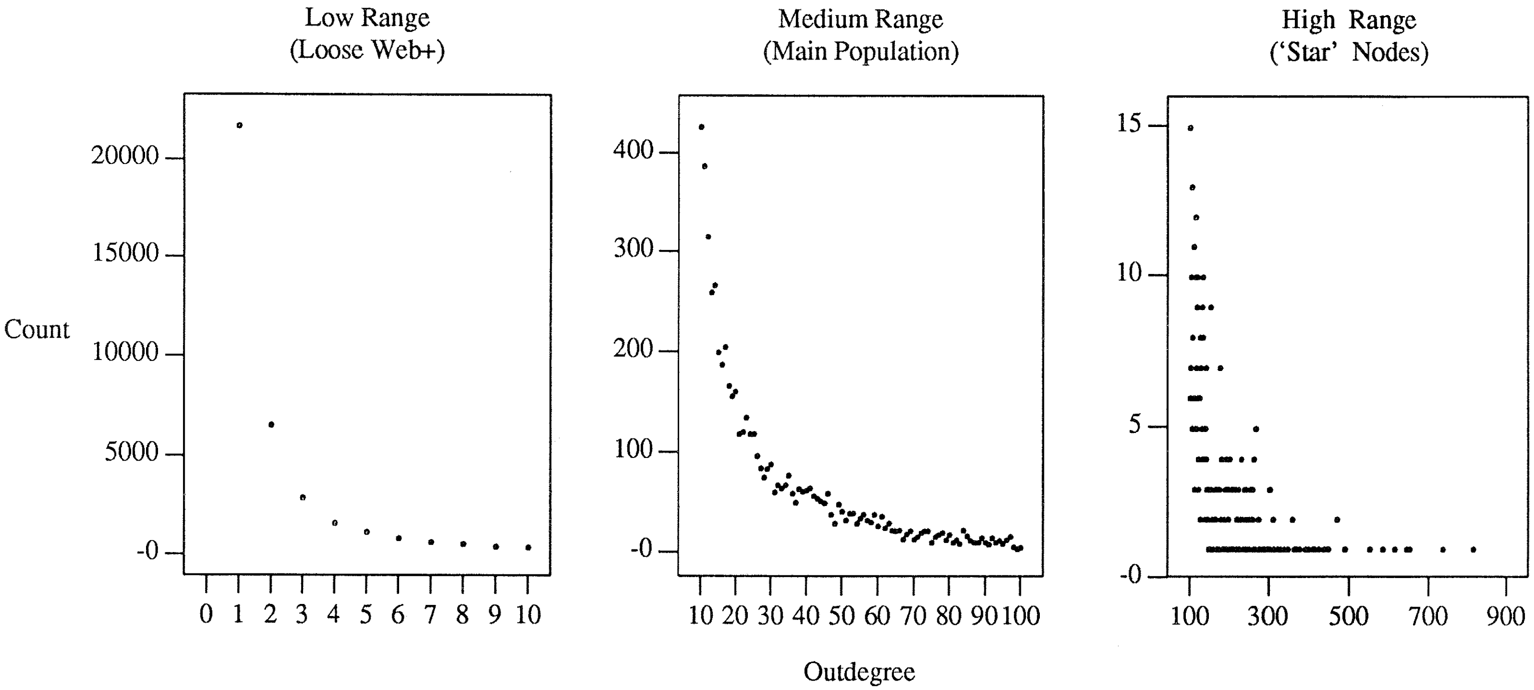

Figure 5: User Level Main Component Outdegree Distribution, In Magnitude Ranges

In this figure, each sub-plot looks exponentially distributed, but with increasingly poor fit towards higher outdegrees, where the counts are small and thus more sensitive to minor variations. The mean outdegree was 7.07, with a standard deviation of 23.35. This small mean reflects the sparse sampling of data from the loose web portion of the graph. Computing the moments on the graph derived by removing the loose web portion of the graph yielded a mean of 45.41, with a standard deviation of 54.87. These figures provide further indication that outside of the loose web, our statistical sample managed to capture a meaningfully large proportion of the true electronic mail communication graph. The low outdegree subplot is labelled "Loose Web+" because the nodes in this plot are actually a superset of the loose web, which consists of all nodes of outdegree 1, plus the nodes of outdegree 2 connected to these nodes, and so on.

Given the range of node outdegrees, we wished to measure the importance of "star" nodes, or nodes with particularly high outdegrees. These nodes represent individuals who are particularly heavily involved in network communication, and hence who might be critical to the structural integrity of the graph. To measure their importance from this perspective, we constructed a graph derived by removing the 564 nodes in the high range outdegree subplot in Figure 5, as well as the edges incident to these nodes. In total, doing this removed 6,597 nodes (15% of the original graph) and 88,242 edges (49% of the original graph). The remaining graph had 178 components, the largest of which is compared with the main component of the User level graph in Table 6.

While the star node-reduced graph's main component is significantly smaller than the original User level graph's main component, Table 6 indicates that its macroscopic structure is remarkably similar to the original User level main component. We also note that 178 components were formed by removing 564 star nodes, for an average of 3.17 star nodes per component disconnection. This measurement and the other measurements in Table 6 indicate that the star nodes are important for improving the interconnection richness of the graph, but not critical to the structural integrity of the graph. This point is interesting, because it indicates a basic level of structural decentralization that provides natural fault tolerance in the communication paths used for electronic collaboration.

The final microscopic level analysis we performed involved studying the individual star nodes. Browsing through the list indicated that most star nodes fell into one of the following classes, in approximately decreasing order of outdegree:

- 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 chairpeople and other managers.

These observations were drawn from the authors' 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 structural integrity of the graph: while the star nodes perform important functions, their functions are mostly independent of the many SSGs that exist for individual shared interests.

In summary, the microscopic graph measurements in this section indicated that nodes have a wide range of outdegrees. Moreover, we found that "star nodes" -- individuals with particularly high outdegree -- were significant in improving the efficiency of interconnection, but not critical to the graph's structure. This observation indicates that the graph has a highly decentralized, naturally redundant structure.

5.3. Intermediate Scope Graph Measurements

In the current section we analyze the graph using intermediate scope measurements, designed to elucidate properties of SSGs within the overall graph.

Ideally, we would like to be able to isolate the particular SSGs to which each node in the graph belongs, and understand how these subgraphs facilitate efficient global electronic collaboration. Efforts to do so are hampered by the fact that the data do not contain any indications of the reasons users communicated. The fact that a set of nodes share a large number of common neighbors is neither necessary nor sufficient to imply that the nodes belong to a common SSG. For example, the subject of their shared communication could be irrelevant administrative information broadcast to all users. Or, two nodes that share neighbors could do so because of shared interests different from those held by other nodes that share these neighbors.

	Nodes	Edges	E/N	Diameter	Sampled Average Path Length (accuracy)	Sampled Maximum Path Length (accuracy)
Main Compon., Original User Level Graph	43,498 (85.6% of full graph)	179,030 (97.4% of full graph)	4.12	21	5.96 (+/- 6.6%)	13.99 (+/- 2.4%)
Main Compon., Star Node Removed User Level Graph	36,261 (98.3% of full star node-removed graph)	90,304 (99.5% of full star node-removed graph)	2.49	22	7.03 (+/- 5.06%)	16.29 (+/- 2.66%)

Table 6: Measurements of Importance of 'Star' Nodes

Unfortunately, acquiring information about the reasons users communicated would have required an unreasonable privacy compromise, and would have posed a difficult natural language recognition problem. Therefore, we chose instead to perform a series of structural analyses of the graph, without such information. The structural analyses are similar in spirit to techniques used for traffic analysis, where communication patterns are used to deduce information about the underlying communication without knowledge of what was being communicated [Callimahos 1989].

Our analysis embodies a notion of an *Aggregate Specialization Subgraph (ASG)*, corresponding to the simultaneous membership of an individual in several closely related principal SSGs, concerning the topics of highest interest to that individual. For example, as will be seen shortly, Schwartz belongs to an ASG concerning networks, distributed systems, privacy/security, and performance. This notion corresponds closely to professional communities, such as those that hold conferences and professional meetings. Any such meeting is attended primarily by individuals who share a number of related interests, i.e., belong to the same ASG.

Our goal was to develop an algorithm that could isolate ASGs within the graph by searching for particularly 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 [Even 1979]. 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 for 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 intuitive sense in terms of the interests known to be possessed by the people in the list.

The first interest distance measure we tried involved computing the number of edges in common between the distinguished node and each node to which it was directly connected. The problem with this measure was that it does not consider the nodes *not* in common between between pairs of nodes. This means that, for example, a node that has a small outdegree but that connects to a number of star nodes would be considered close to any node that was close to those star nodes.

To remedy this problem, we used an interest distance measure defined as the symmetric difference set size over the union set size. In other words, given nodes n_1 and n_2, the distance measure was computed as

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

where $C(n)$ is the set of nodes directly connected to node n. 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 with one another. Note also that the measure decreases as the number of nodes not in the intersection set increases. Using this measure we compute the ASG surrounding a distinguished node as the portion of the computed node list with distance less than 1.0. This portion is always a very small proportion of the total number of nodes in the graph.

Applying this algorithm with Schwartz as the distinguished node, we noticed that the results were muddled by the presence of irrelevant edges generated by local administrative mail. This problem made it appear that the individuals 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. While this type of communication is certainly important from the department's perspective, it typically does not represent interests associated with distributed collaboration. As a simple solution, we performed all intermediate scope graph measurements on the InterDom level graph described at the beginning of Section 5.1, reasoning that interorganizational communication was more likely to exclude this type of administrative communication.

Applying the algorithm using the InterDom level graph yielded better results, providing a list that more closely corresponded to people whose primary professional interests matched Schwartz's. However, the list still contained many erroneous entries, for people whom we knew were not closely related to the distinguished starting nodes. The problem was that nodes with small outdegrees 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 computation described in Section 5.1 (illustrated on the User level graph in Table 5) to derive a core subgraph of the InterDom level graph with 3802 nodes and 9568 edges. We then recomputed the intermediate scope measurements on this new subgraph, and achieved dramatically improved results. As an example of these results, the first 12 entries of computed Schwartz's ASG are described in Table 7. As can be seen, in most cases the individuals isolated by this algorithm had related interests to Schwartz, whose interests lie primarily in networking, distributed systems, privacy/security, and performance. Moreover, several of the individuals in the list were not known personally by Schwartz, indicating that the algorithm managed to uncover relevant individuals with whom Schwartz had never exchanged electronic mail.

Relationship to Schwartz	Distance From Schwartz
Schwartz	0.0000
Graduate student at university where Schwartz studied, involved in networking research	0.7692
Theory professor who studied where Schwartz studied	0.8462
Unknown student at a Southwestern U.S. university	0.8571
Industrial systems researcher who studied where Schwartz studied	0.8824
Schwartz's Ph.D. advisor (interested in performance and distributed systems)	0.9048
System administrator at an East Coast U.S. university	0.9048
Systems and security researcher at a Midwest U.S. university	0.9130
Ph.D. advisor of Schwartz's Ph.D. advisor (interested in performance and systems)	0.9167
Performance and Systems researcher at a West Coast U.S. university	0.9167
System administrator at an East Coast U.S. university	0.9167
Unknown government laboratory researcher	0.9167

Table 7: Top of Computed Aggregate Specialization Subgraph Surrounding Schwartz

Applying the algorithm to a number of other people we knew yielded similarly encouraging results. A particularly interesting example involved a researcher at a university at which we did not collect mail logs. We found that the computed ASG contained a large number of people involved with the Computer Professionals for Social Responsibility. This fact corresponds well with one of that researcher's interests. Runing the algorithm on another person at a university where we did not collect data yielded a list of many persons from the Internet

Activities Board, of which the individual was a member. These examples illustrate the ability of the data collection and analysis processes to correctly characterize the aggregate interests of a much larger population than just the sites where data was collected. This was possible because data collection sites logged traffic bound for many other sites.

Using the ASG isolation algorithm, we computed ASGs surrounding 500 randomly sampled nodes in the core InterDom level graph, to provide statistical characterizations of the ASGs. Figure 6 shows the distribution of ASG size for the core InterDom level graph. As can be seen, the distribution is roughly bimodal. Most people collaborate with only a small number of other people who share their interests, while a small number of people make much more significant use of distributed collaboration. Clearly, the potential exists for significantly more distributed collaboration to take place overall than currently does.

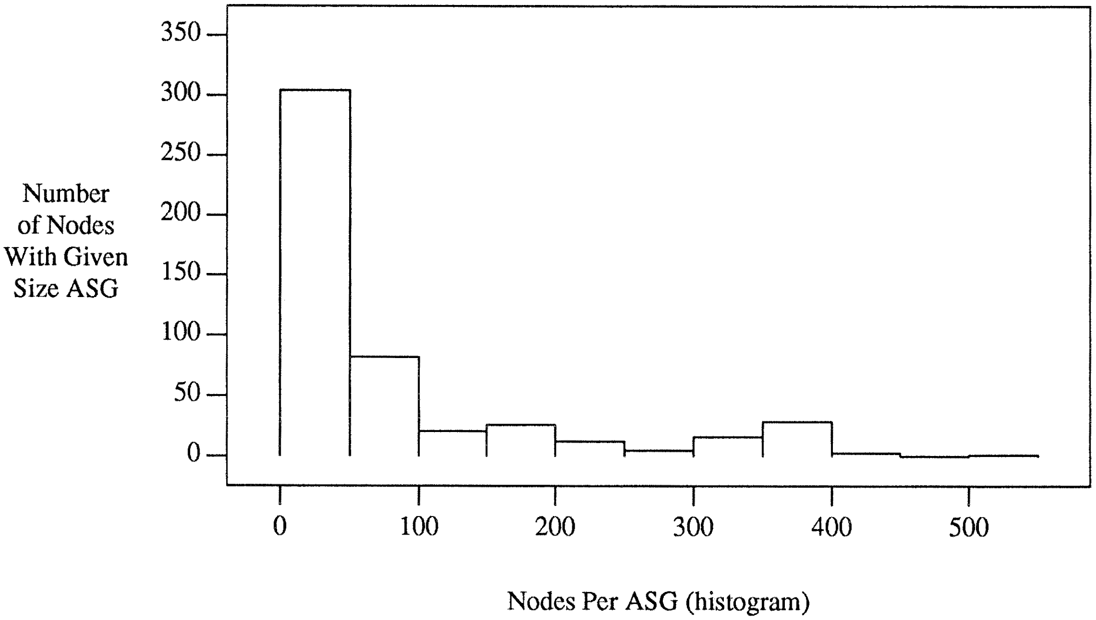

Figure 6: Aggregate Specialization Subgraph Size Distribution

Figure 7 shows the distribution of the number of administrative domains per ASG. Interestingly, although the ASG size distribution is bimodal, the ASG administrative domain distribution roughly approximates an exponential decay. This observation indicates that even for people who collaborate with a fairly large number of individuals, those individuals tend to be clustered at a relatively smaller number of administrative domains.

Figure 8 shows the distribution of the interest distance over the sampled nodes, for nodes other than the distinguished nodes (since distinguished nodes always have distance 0 from themselves). As can be seen, nodes in an ASG tend to be quite distant from the distinguished node. This observation underscores the complexity of the graph, as it indicates that each user tends to communicate with other users who communicate with different sets of users.

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. This is an interesting observation for such a large scale environment, because it indicates that an ASG is like a signature for an individual, in the sense that it represents that individual's specific combination of interests. When interest distances were not included (i.e., just node names were used), 205 (.16%) of the $\left[{500 \atop 2} \right]$ possible pairs were not unique. This point also underscores the fact that users' conversations concerned a large number of different topics. The number certainly exceeds the approximately 600 news groups into which USENET conversations are constrained to fit.

While the ASG surrounding each individual is nearly unique, Figure 9 shows that individuals tend to share closely related interests with a number of other individuals. This figure plots the distribution of average node

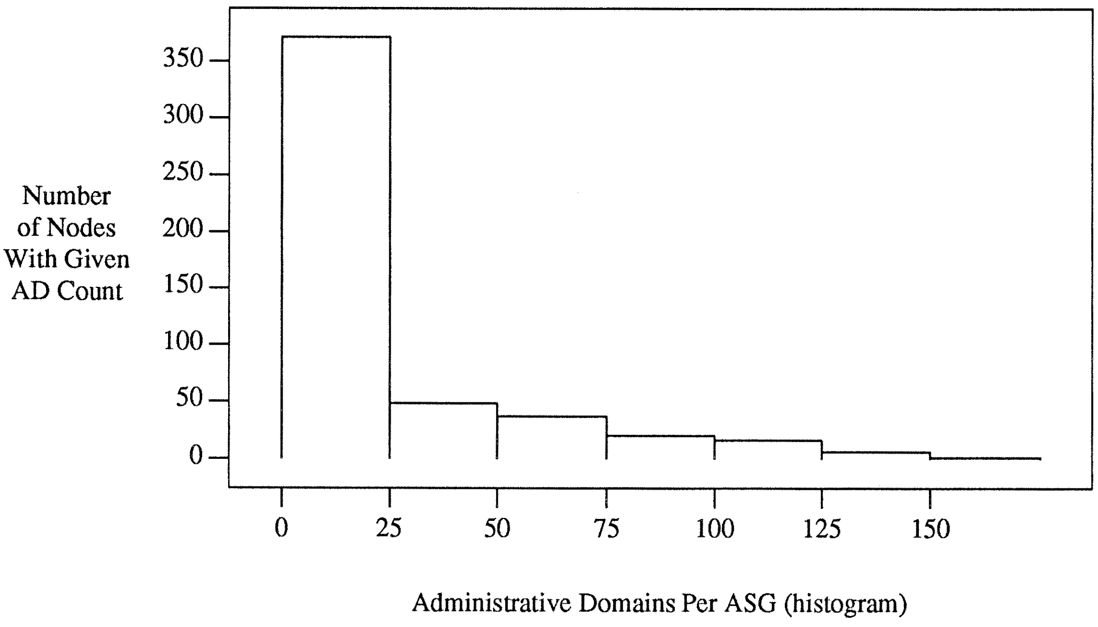

Figure 7: Aggregate Specialization Subgraph Distribution Across Administrative Domains

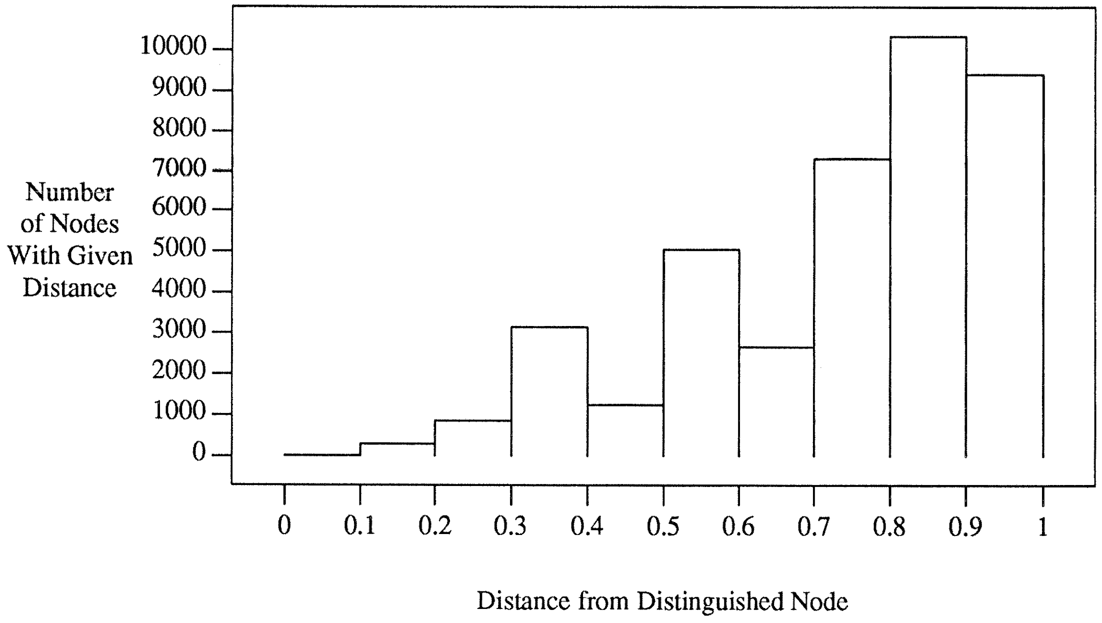

Figure 8: Average Distance Over All Sampled Aggregate Specialization Subgraphs

distances as a function of how many of the closest nodes were sampled. In other words, the Y axis plots the average distance, over all sampled ASGs, of the closest X percent of the nodes in the ASGs, not counting the first nodes (which are the distinguished nodes around which the ASGs were constructed). As can be seen, the nearest 35% of the nodes in ASGs are at a distance of .5 or closer. The curve is concave downward because of

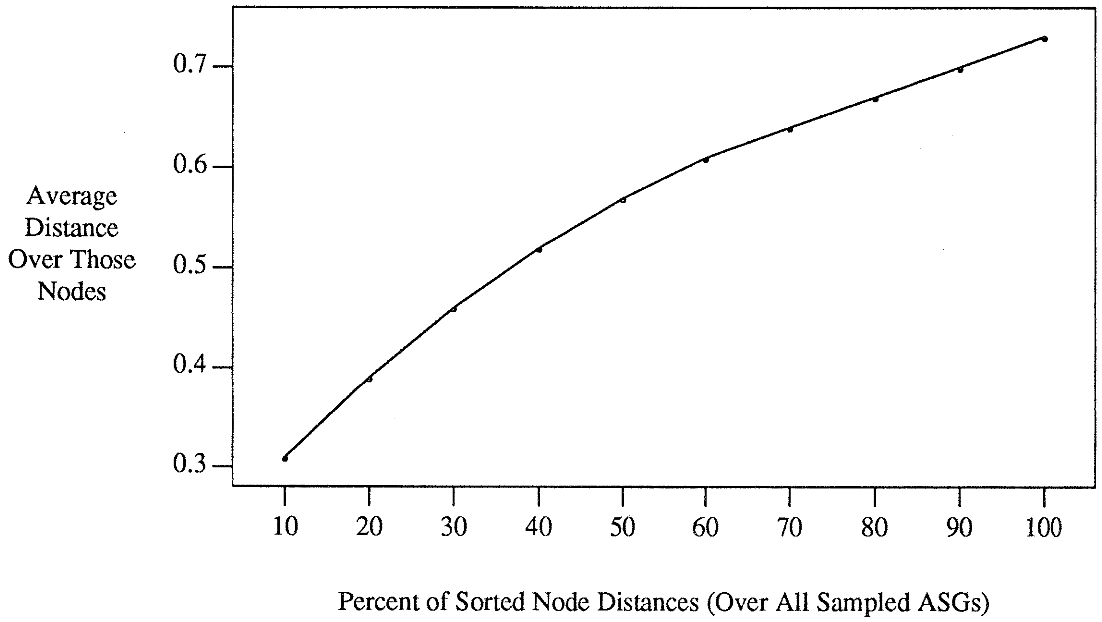

Figure 9: Aggregate Specialization Subgraph Average Distance of Closest Nodes

the relatively large number of nodes that are far away from the distinguished nodes.

6. Potential Applications

The measurements from Section 5 indicate that the electronic mail graph that has naturally redundant, efficient paths for distributed collaboration, yet that distributed collaboration is currently quite underrealized.

An interesting possibility exists to support distributed collaboration using the ASG isolation algorithm we have developed. Rather than specifying sets of users with whom to exchange messages (the electronic mail paradigm) or newsgroups on which to post messages (the news paradigm), a user could simply specify a small set of "seed" people whose interests are believed to be close to a topic of interest. A system could then use the ASG isolation algorithm to form lists of users whose interests are close to those of each of the seed people. The system could then show the user the generated lists and ask the user to select among the entries. This way, shared interests are specified using the communication graph itself, rather than an artificially imposed classification structure. As with bulletin boards, ASG membership disclosure would be voluntary. Individuals who believed that allowing access to their ASG lists posed too severe a potential privacy violation could chose not to participate.

To experiment with this idea, we constructed a computation that generated ASGs around each of a specified set of distinguished nodes, and then summed the quantity $(1-\mathit{distance})$ across the lists, producing a number for each node corresponding to its overall distance from the nodes closest to the distinguished nodes. This *local ASG derivation* produced very interesting lists of persons related to chosen starting nodes, concerning shared interests in such areas as distributed computing, performance analysis, and naming. By specifying only a few seed users, many other relevant individuals were found.

We also built a computation that attempted to produce a global measure of interest distance. Starting from one distinguished node, we generated a list of the closest 100 nodes to that node. Next, we computed the interest distance lists for each of these 100 nodes. For each node in these secondary lists, we summed the quantity $(1-\mathit{distance})$ across the lists, producing a number for each node corresponding to its overall distance from the nodes closest to the distinguished node. The results of this *global ASG derivation* turned out to be quite poor, because each list represents an ASG, and taken in whole no small number of interests would overshadow all other sets of interests. The conclusion to be drawn from this experiment is that it makes most sense to select a small number of carefully-chosen seed individuals in deriving ASGs for distributed collaboration.

Because the average path length between individuals in the overall graph is already quite small, local ASG derivation could likely provide a very effective means to locate a small number of appropriate individuals to contact when searching for or disseminating information, in an automated analog to the resource discovery example illustrated in Figure 1.

Other forms of distributed collaboration could also be supported using local ASG derivation. For example, the technique might be used to organize collections of related information at public archive sites, an application we are beginning to explore [Schwartz et al. 1990]

A final issue raised by this study concerns an important potential privacy sacrifice in electronic mail communication. Using the ASG isolation algorithm we have developed, it is 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 individuals under scrutiny. Given the increasing use of telemarketting techniques and the recent increase in government surveillance in cases like "Operation Sun Devil" [Chapman 1990], this is a dangerous possibility. It would be difficult to protect against these problems, since doing so would require encrypting message headers, which would make mail routing difficult. We believe the solution to this problem lies in legal restrictions that only allow such clustering techniques to be used under explicit consent of the users involved.

7. Conclusions

Hierarchy provides poor support for large scale distributed collaboration. To better understand what organizational support is needed for distributed collaboration, we collected electronic mail logs at 15 universities and companies involved with a variety of research, education, and product development projects. Analyzing the graph structure resulting from these data, we defined and characterized what we have called an Interest Specialization Graph structure. This structure preserves the scalability of a hierarchy by grouping entities into local neighborhoods, without constraining the nesting relationships of these neighborhoods.

The data we collected captured a large snapshot of global electronic mail traffic. Macroscopic graph measurements demonstrated a fairly small diameter and a very small average path length between people. These properties enhance rapid and reliable dissemination of information. Microscopic graph measurements indicated that the number of people with whom people communicate (their outdegree) varies over a wide range. In addition, we found that "star nodes" -- individuals with particularly high outdegree -- were significant in improving the efficiency of interconnection, but not critical to the graph's structure. This observation indicates that the graph has a highly decentralized, naturally redundant structure. This structure is interesting because it provides often-sought properties with considerably more flexibility and less complexity than typical computer systems do.

We next developed a set of algorithms capable of isolating collections of individuals according to shared interests, without knowledge of the content of their communication. The algorithms involve two parts. First, an algorithm was run to produce a subgraph whose organizational structure was separable from the clutter of the statistical sampling. This graph represented communication between a tightly interconnected core group of individuals, communicating across administrative domain boundaries. Second, a clustering mechanism based on a notion of interest distance was applied to particular nodes in the graph, to isolate other nodes with closely related sets of interests. Using these algorithms we measured the intermediate scope structure of distributed collaboration in the global electronic mail community. We found that use of distributed collaboration is split in a bimodal fashion. Most people collaborate with only a small number of other people who share their interests, while a small number of people make much more significant use of distributed collaboration. Also, users tend to share interests with a large number of individuals distributed across a relatively smaller number of administrative domains. Moreover, an individual's aggregate interests are largely unique, even though he/she shares many particular interests with other people, and has closely related interests with a relatively small number of other users.

The shared interest clustering algorithms we developed provide a possibility for organizing distributed collaboration in a fashion different from the primary paradigms currently in use, by dynamically constructing sets of users with shared interests from a small set of specified "seed" users. The potential advantage of this technique is that the notion of shared interests is more dynamic and fine grained than specifying specific users or interest group lists, since it is based on the communication graph itself, rather than an artificially imposed classification scheme. And because the average path length between individuals in the overall graph is quite small, the ASG derivation technique could likely provide a very effective means to locate a small number of individuals to contact when searching for or disseminating information. Other forms of distributed collaboration could also be supported by this technique, including resource discovery among public archive sites.

Another issue raised by the ASG derivation algorithm is the fact that it is possible to deduce shared primary interests simply by monitoring patterns of communication, without access to the text of message traffic. This observation indicates an important potential privacy sacrifice in electronic mail communication that would be difficult to protect against, since doing so would require encrypting message headers, which would make mail routing difficult.

Acknowledgements

We would like to express our appreciation to all of the system administrators around the Internet who allowed us to collect measurements from their systems. We thank Tom Liebert of Merit, Inc., for providing access to the Internet map database needed to construct the geographic map of Section 5.1. We thank Marvin Solomon for suggesting some of the graph computations used in Section 5.1. We thank Andrzej Ehrenfeucht for suggesting the symmetric difference measure used in Section 5.3. We thank Rebecca Marvil and Alan Krantz 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 1.1. Finally, we thank Jonathon Bein, David Goldstein, and David Wagner for their comments on an earlier draft of this paper.

8. References

[Allman 1985]
E. Allman. Sendmail - An Internetwork Mail Router. *UNIX Programmer's Manual, 4.2BSD*, 2C, Comput. Sci. Division, EECS, Univ. of California, Berkeley, June 1985.

[Arms 1989] W. Y. Arms. Electronic Publishing and the Academic Library. Paper presented to the Society for Academic Publishing, 11th Annual Meeting, May 1989.

[Boissevain 1974]
J. Boissevain. *Friends of Friends - Networks, Manipulators, and Coalitions.* Oxford, Blackwell, 1974.

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

[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.

[Conklin 1987]
J. Conklin. Hypertext: A Survey and Introduction. *IEEE Computer Magazine*, 20(9), pp. 17-41, Sep. 1987.

[Donahue & Widom 1986]
J. Donahue and J. Widom. Whiteboards: A Graphical Database Tool. *ACM Trans. Office Information Syst.*, 4(1), pp. 24-41, Jan. 1986.

[Estrin 1985]
D. L. Estrin. Inter-Organizational Networks: Stringing Wires Across Administrative Boundaries. *Computer Networks and ISDN Systems*, 9(4), pp. 281-295, Apr. 1985.

[Even 1979] S. Even. *Graph Algorithms.* Computer Science Press, Rockville, MD, 1979.

[Hoffman 1987]
P. Hoffman. The Man Who Loves Only Numbers. *Atlantic Monthly Magazine*, 260(5), pp. 60-74, Nov. 1987.

[Kahn & Cerf 1988]
R. E. Kahn and V. G. Cerf. *The Digital Library Project - Volume 1: The World of Knowbots.* Corp. for National

Research Initiatives, Mar. 1988.

[Lampson 1986]
B. W. Lampson. Designing a Global Name Service. *Proc. 5th ACM Symp. Principles Distr. Comput.*, pp. 1-10, Aug. 1986.

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

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

[Quarterman & Hoskins 1986]
J. S. Quarterman and J. C. Hoskins. Notable Computer Networks. *Commun. ACM*, 23(10), pp. 932-971, Oct. 1986.

[Schatz & Caplinger 1989]
B. R. Schatz and M. Caplinger. Searching in a Hyperlibrary. *Proc. 5th IEEE Int. Conf. Data Eng.*, pp. 188-197, Feb. 1989.

[Schwartz & Demeure 1989]
M. F. Schwartz and I. M. Demeure. Characterizing Distributed Computing Paradigms. Tech. Rep. CU-CS-422-89, Dept. Comput. Sci., Univ. Colorado, Boulder, CO, Aug. 1989. Submitted for publication.

[Schwartz 1989]
M. F. Schwartz. The Networked Resource Discovery Project. *Proc. IFIP XI World Congress*, pp. 827-832, San Francisco, CA, Aug. 1989.

[Schwartz & Tsirigotis 1990]
M. F. Schwartz and P. G. Tsirigotis. Experience with a Semantically Cognizant Internet White Pages Directory Tool. To appear, *J. Internetworking Research and Experience*, 1(2), Dec. 1990.

[Schwartz et al. 1990]
M. Schwartz, D. Hardy, W. Heinzman and G. Hirschowitz. Supporting Resource Discovery Among Public Internet Archives Using a Spectrum of Information Quality. In preparation, Sep. 1990.

[Sedgewick 1988]
R. Sedgewick. *Algorithms*. Addison Wesley, Reading, MA, 1988. Second Edition.

[Spafford 1988]
E. H. Spafford. The Internet Worm Program: An Analysis. Purdue Tech. Rep. CS-TR-823, Nov. 1988.

[Travers & Milgram 1969]
J. Travers and S. Milgram. An Experimental Study of the Small World Problem. *Sociomety*, 32(4), pp. 425-443, 1969.

[Wulf 1989] W. A. Wulf. Towards a National Collaboratory. Report of Invitational Workshop at the Rockerfeller Univ. Oct. 1989.