Using Rhythms of Relationships to Understand E-Mail Archives

Adam Perer and Ben Shneiderman
Human–Computer Interaction Lab, Institute for Advanced Computer Studies and Department of Computer Science, University of Maryland, College Park, MD 20742. E-mail: {adamp, ben}@umd.edu

Douglas W. Oard
Human–Computer Interaction Lab, Institute for Advanced Computer Studies and College of Information Studies, University of Maryland, College Park, MD 20742. E-mail: oard@umd.edu

Due to e-mail's ubiquitous nature, millions of users are intimate with the technology; however, most users are only familiar with managing their own e-mail, which is an inherently different task from exploring an e-mail archive. Historians and social scientists believe that e-mail archives are important artifacts for understanding the individuals and communities they represent. To understand the conversations evidenced in an archive, context is needed. In this article, we present a new way to gain this necessary context: analyzing the temporal rhythms of social relationships. We provide methods for constructing meaningful rhythms from the e-mail headers by identifying relationships and interpreting their attributes. With these visualization techniques, e-mail archive explorers can uncover insights that may have been otherwise hidden in the archive. We apply our methods to an individual's 15-year e-mail archive, which consists of about 45,000 messages and over 4,000 relationships.

Introduction

Since 1971, e-mail has grown rapidly in popularity and has become a central part of many users' personal and professional lives. Despite its impressive role in society, there are still few tools available to explore archives of e-mail. The need for such tools will grow as valuable e-mail archives increase in availability. The U.S. National Archives preserves e-mail as government records (Baron, 1999), a recently released collection of Enron e-mail has attracted significant public attention (Grieve, 2003), and some individuals have now accumulated e-mail collections that span decades. Historians and social scientists will undoubtedly find these archives to be a valuable basis for understanding the individuals and organizations that created them; however, it is currently far from clear how these explorers will gain the context they need to understand the archive's numerous conversations.

Table 1 illustrates one way in which the universe of tools for interacting with online conversations can be subdivided. E-mail is created by individuals, and often in some organizational or social context. There has been a great deal of work on individual and organizational e-mail productivity tools (Regions A and B), and on the management and analysis of conversations in public e-mail venues such as mailing lists and Usenet News (Regions C and F). Our work in this article focuses on Region D, as we present new techniques for exploring the archived e-mail of an individual.

Although the principal content of e-mail is free text, when attempting to browse archives, the shortcomings of a text-only display become clear. E-mail archive explorers have previously tackled the archives by keyword searching, but this approach often will result in losing a conversation's context (Donath, 2004). Visualizations are one way to provide this missing context. In this article, we show that valuable information can be uncovered by visualizing the temporal rhythms of social relationships that are evidenced in e-mail archives. Each relationship that is evidenced in an e-mail archive has a rhythm that can be characterized by the intensity of the correspondence over time. Relationships that are brief but intense have rhythms with sharp growth and steep decline. Relationships that are durable and strong have consistent and continuing rhythms. This article presents insights achieved by analyzing the rhythms, which help archive explorers question why certain relationships start and stop, why certain relationships share similar activity patterns, and the nature of the relationships that yield different interaction patterns.

Detecting long-term rhythms, our focus in this article, requires a collection spanning many years. Ben Shneiderman, a coauthor of this article and a pioneer in the fields of

Received May 8, 2005; revised November 29, 2005; accepted December 1, 2005

 ● Published online 2 October 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20387

TABLE 1. E-mail management task decomposition.

	Individual	Organizational	Social
Current	Region A: *Managing an individual user's current inbox*	Region B: *Managing current e-mail within an organization*	Region C: *Managing current conversations in a public space*
Archived	Region D: *Exploring an archive of an individual's messages*	Region E: *Exploring an archive of an organization's messages*	Region F: *Exploring an archive of a public space*

Human–Computer Interaction (HCI) and Information Visualization, has archived the e-mail he produced and received since 1984. The archive contains over 4,000 of Shneiderman's relationships, totaling around 45,000 messages. That archive spans a longer period than any other collection that was available to us when we started this work, offering us a unique opportunity to study the long-term rhythms of relationships present in a real e-mail collection. In the next section, we review related work on interacting with online conversations. Next, we define what we mean by "relationships" and the "rhythms" that they produce. We then present our analysis methods and illustrate the use of those methods on the Shneiderman archive. Finally, we conclude with some suggestions for future work.

Previous Work

In this section, we briefly review prior work on e-mail management, organizing the discussion using the task decomposition shown in Table 1. Interaction with the user's own current e-mail (Region A) is by far the most actively studied e-mail management task in the research literature. An early ethnographic study by Mackay (1988) provided compelling evidence that different people deal with large quantities of their personal e-mail in many different ways. Whittaker and Sidner's (1996) later study resulted in the same conclusion while also describing tasks that individuals use e-mail for beyond the asynchronous communication for which it was designed. Recent attempts to integrate visualizations into e-mail clients seek to help users better manage their e-mail. For example, enabling users to see the thread structure provides them with a better understanding of how conversations evolve over time (Kerr, 2003; Venolia & Neustaedter, 2003). Another example is the Remail project, which provides a "correspondents' map" that allows users to quickly see who they have not replied to recently as well as a "message map" to see messages with similar attributes (Rohall et al., 2003).

Some recent projects have investigated exploration of personal e-mail archives to uncover trends and patterns (Region D). PostHistory explored e-mail archives that extend as far back as 5 years, seeking to support the development of insights that would be socially relevant to the owner of the e-mail (Viegas, Boyd, Nguyen, Potter, & Donath, 2004). PostHistory featured an interface that animates over time to allow users to get a sense for their steady and intense relationships, and to illustrate fast-paced rhythms (e.g., resulting from project deadlines) and slower-paced rhythms (e.g., during vacations). Social Network Fragments, by contrast, focused on revealing groups of correspondents that emerge through e-mail exchanges (Viegas et al., 2004). This interface also used time as a dimension to see how connections among correspondents appear and dissolve, thereby providing a way for the user to visualize the evolution of their own social network. In small studies, users were able to see meaningful patterns with both PostHistory and Social Network Fragments, sometimes using the visualization as instigation for telling stories.

The ubiquity and persistence of e-mail has important consequences for the management of information within organizations (Region B). Ducheneaut and Bellotti (2001) studied the use of e-mail in three organizations and discovered that patterns of e-mail use vary with individual roles within those organizations. They also noted that characteristics of each organization influenced the ways in which people used and organized their e-mail collections. Tyler and Tang (2003) added to the understanding of e-mail use within organizations, observing that responsiveness patterns vary in ways that reflect the dynamics of interpersonal relationships within an organization. That observation led them to suggest that tools for estimating expected response latency could help users detect communication breakdowns. Another example of an organization tool is the "Email Mining Toolkit," developed by Li, Hershkop, and Stolfo (2004) to support anomaly detection by creating behavior models. They then used these models to detect aberrant behavior of individuals or groups that may indicate abuse or policy violations.

Exploration of archived collections of organizational e-mail also has been studied (Region E). Tyler et al. (2003) used the social network analysis concept of "betweenness centrality" to identify communities in a large collection of e-mail from a single organization, discovering that evidence of the management hierarchy for that organization could be found in the structure of the resulting graph. Leuski et al.'s (2003) "eArchivarius" system combined clustering based on content or co-addressing with activity timelines and biographies to explore activities in the U.S. National Security Council during the Reagan era using a small collection of declassified e-mail messages.

Usenet News, a distributed management system for a large collection of public mailing lists, has been archived since 1981. Mailing list usage differs somewhat from the use of personal e-mail, both because privacy expectations are reduced and because the group-oriented communication structure alters interaction patterns. Smith (1999, 2002) used the "NetScan" system to study social accounting metrics for Usenet participation (Region F) and reported statistics on authorship and on activity over time. Usenet News is immediately available to both participants and nonparticipants ("lurkers"), which makes the distinction between management

DOI: 10.1002/asi

and exploration somewhat less defined than it is in the case of individual and organizational e-mail. Users of the NetScan system can, for example, use it to find intense discussions and related "newsgroups" (Region C). Sack's "Conversation Map" (2000) also explored Region C, focusing on the structure of long-term conversations by using social network diagrams, lists of discussion themes, and semantic network representations to support visualization of conversational structure and content.

The work described in this section is, of course, only a small sample of the extensive research on e-mail utilization that has been reported since the first e-mail was sent over the ARPANET in 1971. Looking broadly at that body of work, however, two trends emerge. First, the vast majority of the reported research has focused on managing current activities rather than on understanding what happened in the past. There has been much less work done in Regions D and E. That makes sense since only recently has e-mail's ubiquity become clear, and archives of e-mail are accruing. Second, the retrospective analyses on individual e-mail (as opposed to mailing lists or Usenet News) that have been done have had limited scope; we are aware of only one study that has looked at even 5 years of e-mail. In this article, we take a longer view, looking back at a 15-year period that spans 1984 to 1998, as Internet e-mail moved from adolescence to adulthood.

Relationships in E-mail Archives

In this section, we describe the e-mail collection with which we worked and the analytical framework that we applied to explore the long-term rhythm of relationships in that collection.

Shneiderman Archive

This archive begins in 1984, 1 year after Ben Shneiderman received tenure as an Associate Professor and founded the Human–Computer Interaction Lab at the University of Maryland. We chose to limit our study to the first 15 years, culminating in 1998, because Shneiderman changed his e-mail file structure significantly in 1999. The resulting set includes 44,971 messages. That is certainly not every e-mail received or sent by Shneiderman during that period. Rather, it includes those that Shneiderman purposefully stored. Although analysis of the results of intentional retention will not provide a complete picture of an individual's e-mail traffic, it does serve to filter out spam and other less significant messages. The saved e-mail gives historians a picture of what Shneiderman felt at the time were the significant conversations in his professional life; however, our analysis will miss some subtle and friendly exchanges which also could serve as sources of interesting rhythms (e.g., as described by Tyler & Tang, 2003).

Relationships

E-mail provides a medium in which users may foster relationships with individuals, organizations, and a global community. Relationships are fundamental to any form of human interaction, so we have chosen to aggregate this collection by relationship rather than the more commonly studied granularities of "threads" (i.e., reply chains) or individual messages. Aggregation into relationships facilitates exploration by masking some sources of variation (e.g., multiple e-mail addresses for a single individual or individuals that participate in multiple relationships) that might otherwise conceal the broad themes that we wish to uncover. By "relationships," we mean a set of conversations over time that reflects a type of interaction that was meaningful to the person who created the e-mail archive. Examples could include conversations with a specific colleague, discussion of a particular topic (e.g., academic governance) involving several members of an organization, or a group of messages regarding the planning of an event (e.g., a professional conference).

The process of discovering unique identities in an e-mail archive is not trivial, especially when dealing with an archive than spans 15 years. People move to various organizations and universities, obtain new e-mail addresses, change their surnames, and evolve their academic interests. For this reason, individuals are not classified simply based on their e-mail header information. Instead, each relationship is identified with help from Shneiderman's filing metadata, as he typically stored relationships in separate folders. Conversations with individuals were usually stored in a folder labeled with their name. Conversations occurring with many participants on a particular topic, such as organizing a conference, were usually stored in a folder labeled with a description of the topic.

We were interested in applying our techniques to learn about Shneiderman's professional life, not his personal life. In the archive, several relationships were present that did not include any content related to his professional career. These relationships include his family, and friends from outside his professional circle. Only 19 of the 4,051 relationships in his archive fell under this category, resulting in a small number of deletions. Those relationships were manually tagged and deleted before any analysis was performed.

To take advantage of the manually tagged relationships, a significant amount of work was necessary to ensure that the data's representation was valid. Occasional misspellings were present, surname ambiguities occurred over time (e.g., folders named "norman" in early years versus folders named "normandon" and "normankent" in later years), and an occasional misstep from naming conventions (storing a message from Catherine Plaisant in a folder named "catherine" instead of "plaisant"). These findings are consistent with those of Ducheaneut and Bellotti (2001), who remarked about users' confusion about storing a message from a corporate colleague in a folder named after the company or the person. These inconsistencies were corrected by fixing typographical errors and standardizing the naming convention for relationships that contained conversations with similar e-mail addresses.

Before our analysis, Shneiderman categorized each relationship into one of three distinct groups. A relationship could be tagged as a *person*, which meant the messages in

DOI: 10.1002/asi

that folder revolved around the relationship of a single person. A relationship also could be tagged as an *organization*, which meant the messages contained within that folder revolved around a variety of individuals communicating about or within the same organization. Finally, the relationship could be tagged as a *topic*, which meant a variety of people from one or more organizations communicating about a similar topic. Of the 4,051 relationships, almost 95% were tagged as people (3,836), compared to only 197 organization relationships and 18 topics.

Note that our human-assisted categorization methods are not a strict requirement for exploring archives. For example, relationships could be postulated automatically based on e-mail addresses and/or message content; however, the availability of Shneiderman's personal categorization scheme gave us comfort that we would be analyzing an accurate representation of the corpus, reducing the "noise" present in our rhythms.

Rhythms of Relationships

By the "rhythm of a relationship," we mean the pattern of activity for a relationship over the duration of an e-mail archive. For example, in Figure 1, two relationship rhythms are shown. The left rhythm depicts a relationship that was inactive during the early years, becomes active in the middle years, and then grew to be an intense relationship in the later years. Conversely, the rhythm on the right shows a relationship that

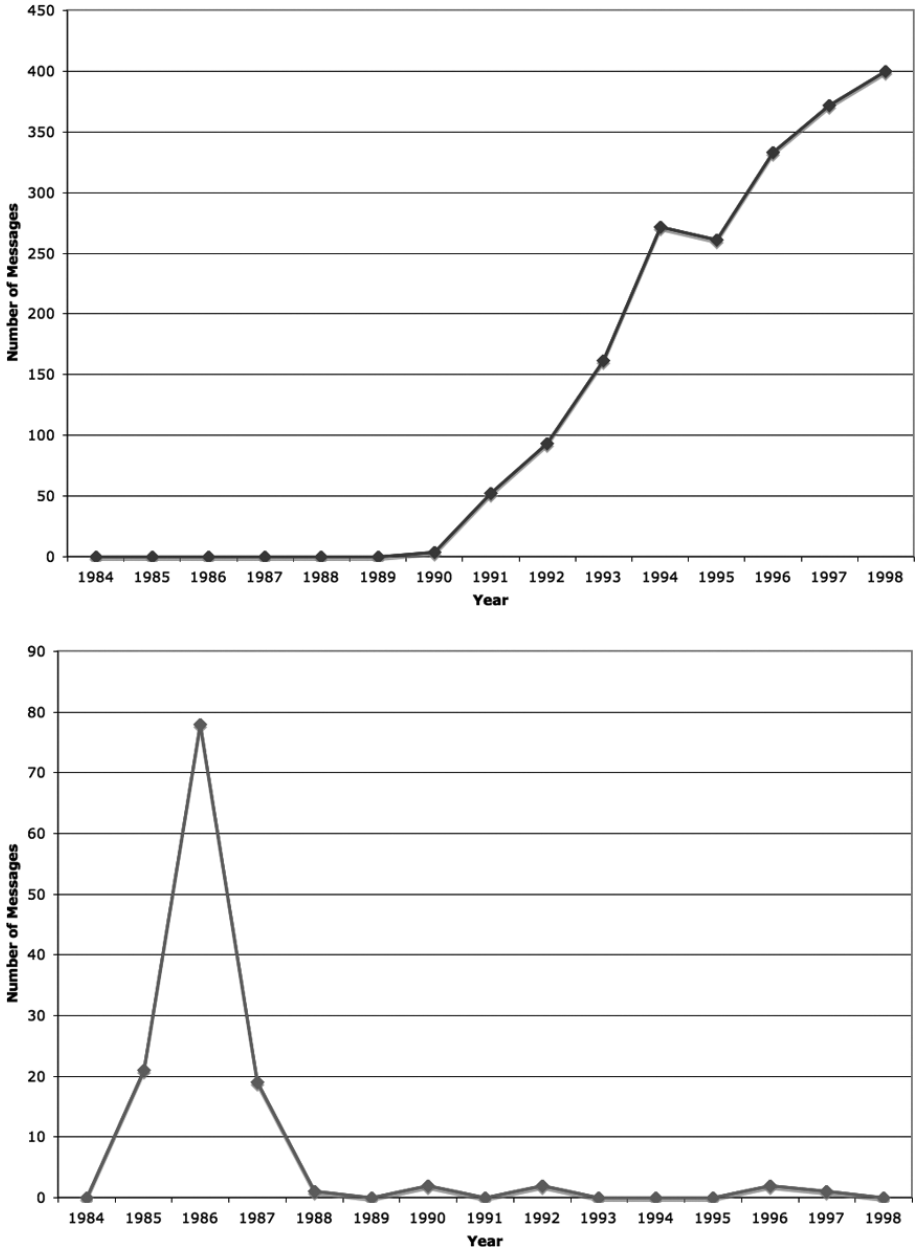

FIG. 1. Examples of two types of rhythms of relationships. On the top, e-mail contact begins only in 1990 and grows steadily, while on the bottom, e-mail contact is intense during 1985–1987 followed by episodic contacts.

DOI: 10.1002/asi

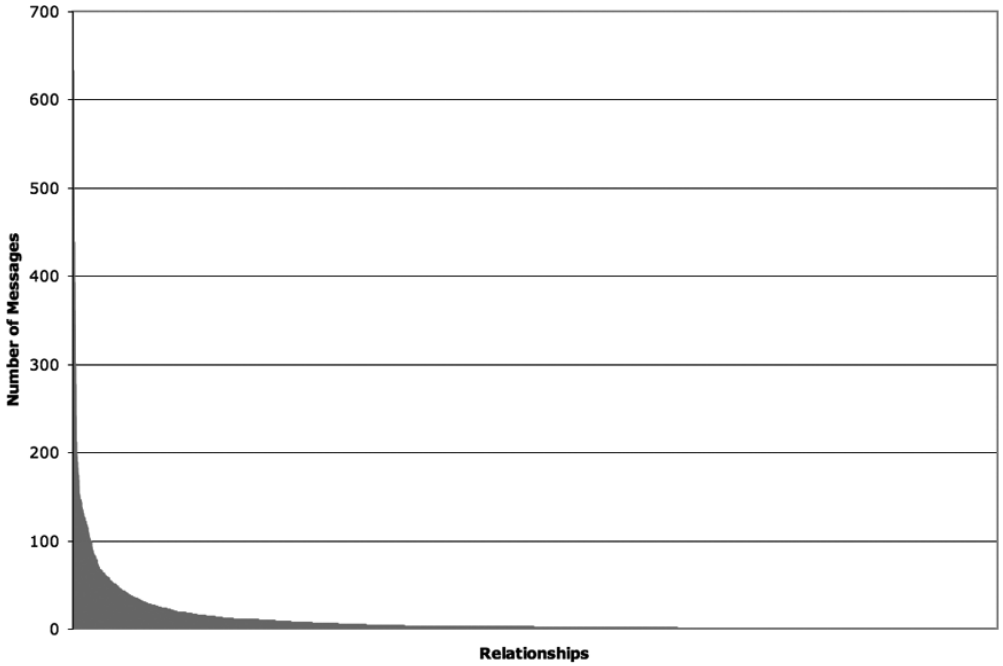

FIG. 2. Distribution of relationships reveals a highly skewed pattern with only 76 relationships having more than 100 messages, and the vast majority of relationships contain just a few messages.

starts out intensely and then eventually becomes sporadic contact. These types of rhythms can be extracted from information that is present in e-mail headers alone, thereby minimizing the need for access to text in the bodies of the e-mail that would naturally be more problematic from a privacy perspective. Due to our interest in understanding long-term patterns, we construct rhythms that have a granularity of a year. Kleinberg (2002) identified similarly bursty patterns over shorter time scales in his own e-mail, suggesting that emerging topics might explain some of this effect.

The rhythms of relationships featured in this article were constructed by writing computer programs that processed the e-mail archives. Our processing tools were written in Java (about 5,000 lines of code) and identified the characteristics of each e-mail message in the archive, such as the sender, recipients, subject lines, date, folder path, and body. These fields were then indexed into a format understood by Lucene, an open-source text search engine. Queries then were run over the index to retrieve counts of messages for the senders/-recipients and folder paths applicable for each relationship identity. Finally, we filter these counts into annual buckets, which assemble to form a rhythm that spans the duration of the archive.

Profiles of Shneiderman's Most Active Relationships

Clearly, not all relationships are made equal; certain relationships are very intense whereas others are quiet and infrequent. In fact, about one third (31%) of relationships in the Shneiderman archive have less than two messages, and 55% have less than four messages. Only 11% of the relationships present in the e-mail archive ever reach 20 or more messages.

Examining the key relationships in an e-mail archive provides an understanding of the nature of the owner's work. Since the Shneiderman archive consists of just 3,836 individual relationships, it is likely that the contents are tied to only the most valued relationships. To gain an understanding of the most frequent correspondents, we extracted the relationships with 100 or more saved messages, leaving only 76 professional relationships.

These 76 professional relationships were just 2% of the 3,836 professional relationships, but they produced 12,771 saved messages (31%) of the 41,420 saved messages. The distribution of relationships is seen in Figure 2. We expect this distribution to be common in e-mail archives of individuals, with a bulk of the messages tied to a small number of key relationships.

Having the archive's owner as a coauthor is not a luxury we expect most historians and social scientists to have; however, we exploited this opportunity to obtain accounts of the identity of these 76 most active relationships. This knowledge is useful, as we can judge our techniques against these verifiable truths. The information provided by Shneiderman is described next, as it provides insight into the types of intense relationships that emerge in a 15-year email archive.

The top 10 most active professional relationships had between 240 and 634 total messages. These relationships included four key colleagues at the University of Maryland (Plaisant, Marchionini, Norman, Chimera), conference organizing partners (Light, Soloway, Rotenberg), and collaborators on other projects (Simons, Ahlberg, Grudin). These reflect Shneiderman's major projects, some with a small number of intense years of activity with over 140 saved

DOI: 10.1002/asi

TABLE 2. Shneiderman's most active relationships, categorized by role.

Most Active Professional Relationships more than 100 saved messages (n = 76)	No.	Average Years Active	Average Total Messages
UMD–Close colleagues	11	9.2	209.7
UMD–Superiors and staff	11	9.6	123.0
UMD–Students	9	9.0	183.8
Colleagues at other universities	17	11.3	152.4
Conference partners	10	8.3	172.7
Corporate partners	9	9.1	137.6
USACM Public Policy	4	5.5	252.3
Book editorial workers	3	8.7	183.0
Government partners	2	9.5	171.5

messages (Ahlberg, Simons, Light, Rotenberg) while the rest showed a more steady pace of exchanges.

These 76 most active relationships were relatively easy for Shneiderman to assign to categories (Table 2). On a large table, he created a small card for each relationship and sorted them into clusters. About a dozen of the names had more than one role, such as when a University of Maryland colleague moved to another university, a former student became a corporate partner, or a book editorial worker also was a colleague at another university. Assignment was by major role, as determined by the majority of saved messages rather than duration.

As expected, many of the most active professional relationships were from the University of Maryland, with 11 being close colleagues, 9 being students, and 11 others being superiors (chairs, deans) and staff (secretaries, administrators). Colleagues at other universities accounted for 17 of the most active professional relationships while conference organizing partners and related efforts covered 10 relationships. Corporate partners including financial supporters, consultancies, and book or lecture collaborators covered 9 relationships.

Other important relationships included 4 colleagues tied to the USACM Public Policy group, in which Shneiderman was a member of the Executive Committee. Development of Shneiderman's book, *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (Addison-Wesley Publishers), showed strong activity for 3 people in the years when the first edition (1986), second edition (1991), and third edition (1997) were in production. Finally, close collaboration with 2 government partners at the National Library of Medicine and the Library of Congress generated high levels of activity for several years.

Methods for Understanding E-Mail Archives

In this section, we identify certain tasks that lead to insights by analyzing the rhythm of relationships in e-mail archives. For each task, we describe the visualization methods that led to the insights and the set of features on which that visualization was based. We illustrate the utility of these analysis methods with examples from the Shneiderman archive.

Evolution of Relationships

With a corpus that spans 15 years, it is to be expected that the nature of some relationships will change over that period. By examining relationships individually, it is possible to witness certain relationships blossoming while other relationships conclude. However, when looking at all the relationships together, one might wonder what sorts of collective patterns emerge: Did the frequency of archived e-mail change as e-mail became more ubiquitous? Are there specific periods in time when the social circle changed more rapidly? Questions of this type can be answered with the following approach.

One of the simplest analyses that can be done is to count the number of messages over time. Figure 3 illustrates the rapid growth in the number of archived messages over time, increasing from 98 e-mail messages in 1984 to 8,499 in

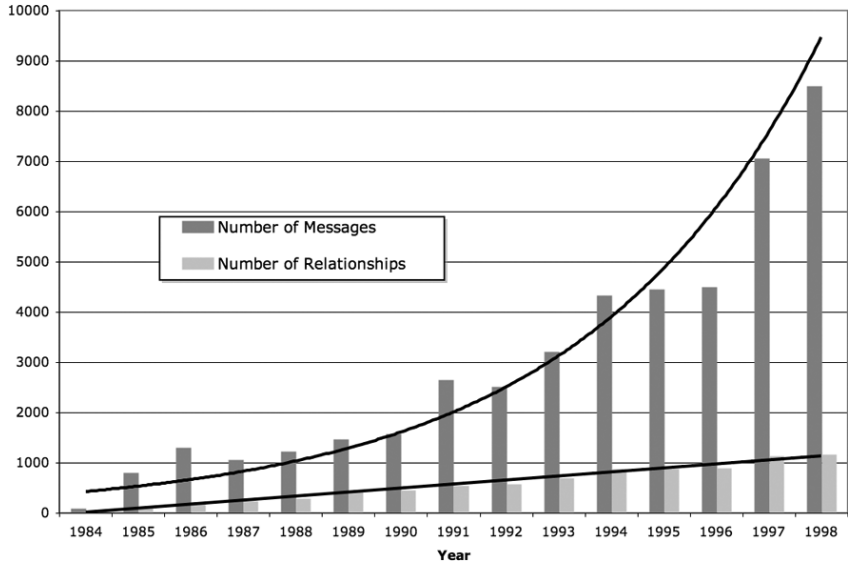

FIG. 3. Growth rates for messages is close to linear over the time studied while the number of messages grows almost quadratically.

DOI: 10.1002/asi

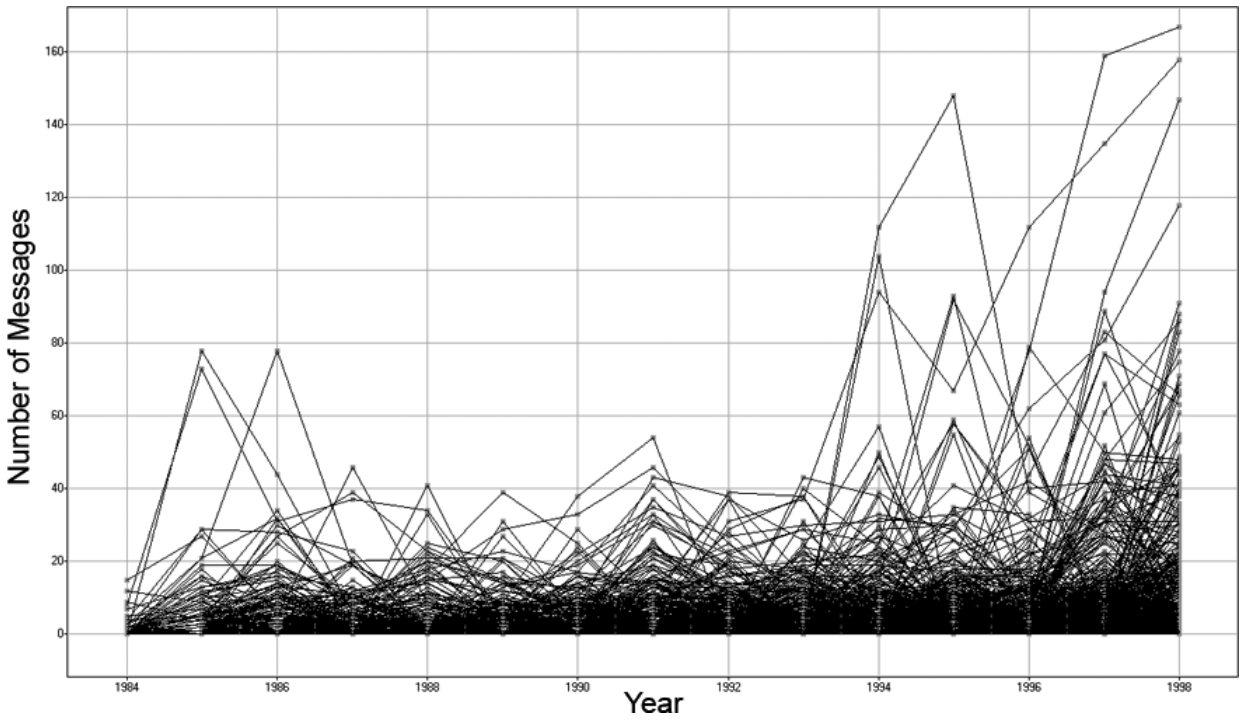

FIG. 4. Among the 4,000 relationship rhythms, interesting features include the three rhythms that have early peaks, and the four relationships with more than 100 messages in 1995–1998.

1998. Figure 3 also shows the number of active relationships, counted for each year over the same period. The growth in the number of active relationships is well fit by linear interpolation, and the growth in the total number of messages is better fit by a quadratic function than by an exponential function. This archive spans a period in which the number of ARPANET/Internet users grew exponentially, and in that context, the more sedate linear growth in the number of relationships is interesting.

By counting the number of messages and active relationships over time, explorers can get a sense of how an e-mail archive evolves. Interesting characteristics can be determined, such as if the individual fosters more relationships over time and if the growth is consistent with the growth of the Internet. The limitations to this approach are that these averages mask considerable individual variation, witnessed in Figure 4, which provides a superimposed image of over 4,000 relationship rhythms from the archive. Figure 4 also illustrates a somewhat surprising (and presently unexplained) absence of brief-but-very-intense relationships during the middle years of the archive.

Relationship Rhythm Patterns

Useful insights about relationships can be discovered based on the pattern of its rhythm. For example, if a historian was looking for evidence of relationships that were strongly related to a temporal event, a search tool that could find relationships that peaked around the time of the event might be useful. One way to support this is by allowing the user to sketch a graph to query the time series, a technique introduced by Wattenberg (2001).

Figure 5 illustrates an example of this type of search on the Shneiderman Archive using the "Hierarchical Clustering Explorer" (HCE; Seo & Shneiderman, 2002). Suppose the searcher postulated that Shneiderman's activities related to policy issues grew markedly in the mid-1990s. If they had an interest in exploring relationships that were unique to that period, they might then construct a query (represented in Figure 5 by a bold line), seeking relationships that sharply grew in 1994, peaked in 1995, and declined in 1996. Rhythms that match this query are shown as thinner lines. The gray background provides a contour based on most active relationships in the corpus for each year. This technique allows explorers to quickly find relationships that follow expected patterns. Of course, there also are situations in which a searcher may not have a specific question in mind when beginning to explore an archive. In this case, providing the searcher with clusters of similar rhythms might offer a point of departure for further investigation.

K-Means Clustering

Clustering based on similarity can be a useful way of revealing characteristic rhythms. Figure 6 shows the result of clustering the 76 most active relationships (i.e., those with the largest total number of messages) in the Shneiderman Archive into nine clusters. We applied *k*-means clustering (MacQueen, 1967) to the 15-year rhythms of these active relationships using Spotfire DecisionSite (Spotfire, 2005).

DOI: 10.1002/asi

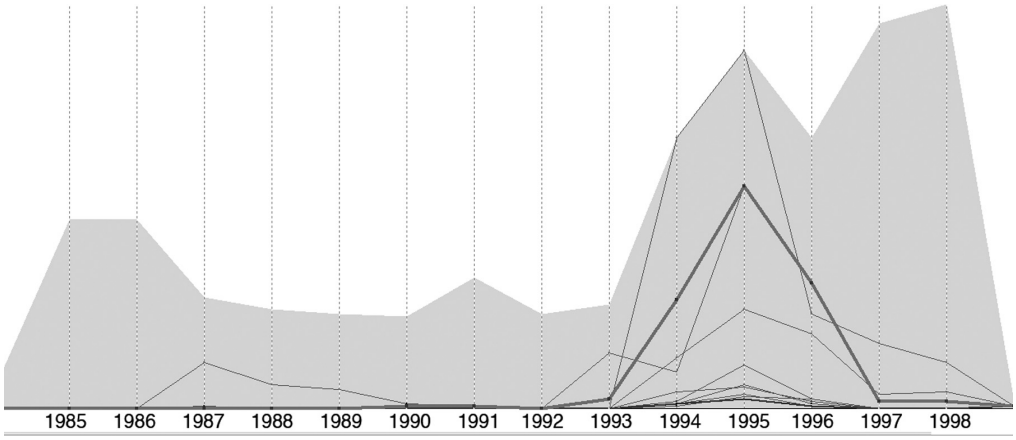

FIG. 5. Search an e-mail archive with a rhythm query. The bold line is the search pattern, and the close matches are shown as lighter lines.

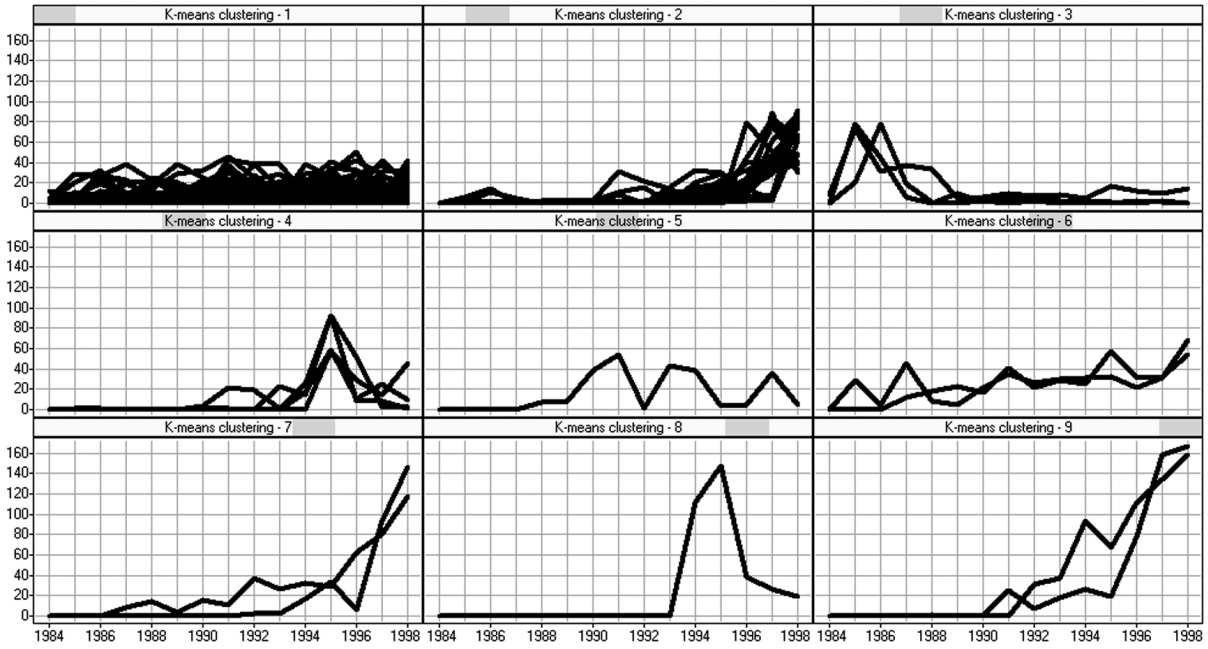

FIG. 6. Nine groups found using *k*-means time series clustering on the 76 most active relationships. The most common pattern (1) is a steady rhythm, and the second most common (2) has a rising rhythm. Interesting clusters are (8) which has a large peak in 1995 and (7 + 9) which show strongly rising patterns.

The number of clusters, *k*, is a parameter of the algorithm. The *k*-means algorithm then divides the 76 rhythms into *k* clusters until the total Euclidean distance between the rhythms and their cluster's centroid is minimized.

Choosing an appropriate *k* is a difficult choice, especially for a searcher unfamiliar with the overall structure of the rhythms or archive. In our initial run, we asked the archive's owner, Shneiderman, to group every relationship with more than 100 messages into distinct groups. By printing out the names on cards and sorting the 76 relationships manually, he came up with the nine distinct groups listed earlier in Table 2. Note that these categories were not chosen based on rhythm patterns. Rather, groups were chosen based on the roles of the people (e.g., academic colleague, corporate collaborator, or graduate student). There was no evidence that each of these roles should constitute their own rhythm clusters, but it provided an interesting value of *k* to start with.

The *k*-means clustering algorithm provides meaningful results, as it successfully displays similar patterns such as those that accelerate in the later years (Cluster 2), relationships that start strong and then die down (Cluster 3), and relationships that peak in similar years (Cluster 4); however, this algorithm classifies most of the relationships into the first cluster, providing little useful information on that set. Selection of a different number of clusters might yield more insight in those cases, but in general, users often find a priori selection of the number of desired clusters to be problematic. In addition, the clusters found had no noticeable correlation with the clusters identified by Shneiderman in Table 2.

Hierarchical Clustering

Hierarchical clustering is another algorithm that can group similar rhythms, but does not require a predetermined

DOI: 10.1002/asi

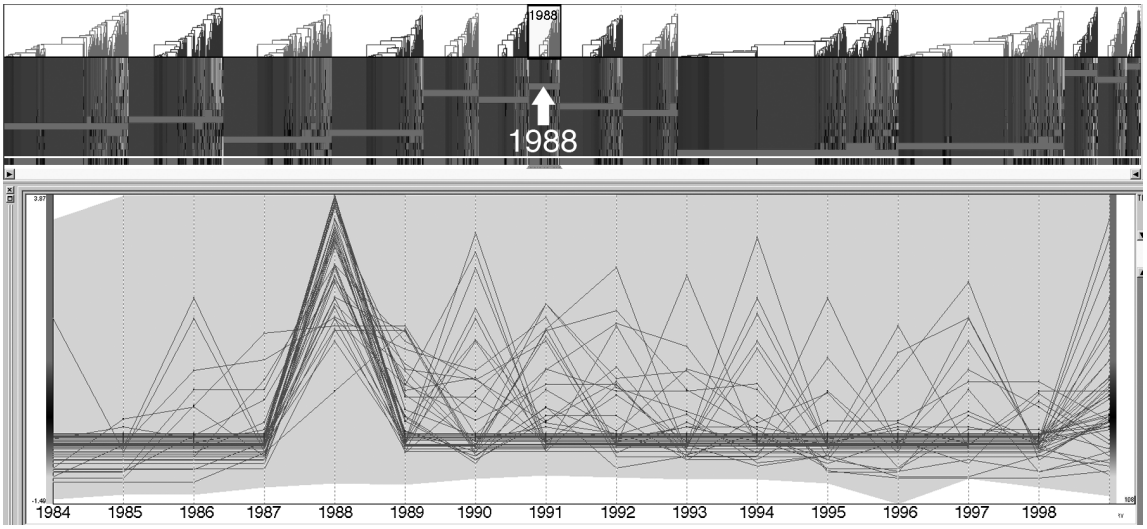

FIG. 7. Hierarchical clustering results on all 4,051 relationships. The cluster with strong activity in 1988 has been selected in the upper panel and is shown in detail on the lower panel. Some of the relationships with strong activity in 1988 had strong activity in other years.

number of clusters (Johnson, 1967). Hierarchical clustering works by finding the pair of relationships with the most similar rhythms, where similarity is computed based on the Euclidean distance between points in the vector space defined by the sequence of annual message counts. It then iteratively builds a hierarchy by pairing these relationships with each other or with an existing cluster of similar relationships. Figure 7 shows results of hierarchical clustering using HCE on all 4,051 relationships. The hierarchy that HCE builds is shown using a dendrogram, displayed in the top panel of the figure. Each subtree of the dendrogram, alternating in gray and black, represents the cluster of relationships that were most intense in each of the 15 years. These subtrees are not arranged in chronological order, but instead retain their order from the constructed dendrogram. These subtrees lead down to the leaves, where each relationship is represented as a column of tiles. Each tile in the column is shaded to correspond to that relationship's intensity in a given year. In this figure, gray shading means a strong intensity.

The subtree surrounded by a black box at the top, labeled "1988" and in the middle of the dendrogram, represents those relationships that were most intense in 1988. Notice how the tiles below this subtree have an obvious gray line in the fifth row of the columns (We annotated this row with a white arrow for clarity.) That row represents 1988, and the shading conveys the large number of messages. The rhythm profiles that correspond to the selected subtree are shown in the bottom panel, where the intense activity in 1988 among these relationships is confirmed.

Hierarchical clustering also detects groups of relationships that are similar beyond 1 year. Subtrees of the dendrogram isolate relationships that have peaks in multiple years. For example, the algorithm constructs a subtree for those relationships that have modest intensity in 1996, grow a great deal in 1997, and then grow a little more in 1998. Looking at this cluster's list of relationships, the four most intense relationships involving Shneiderman's interest in policy are found (Gelman, Brownstein, Ellis, & Simons). This provides evidence that clusters can convey meaning, as the four relationships, remarkably, can be identified when using HCE to zoom in on the subtree (as shown in Figure 8, a view which shows only 2% of the entire tree structure);

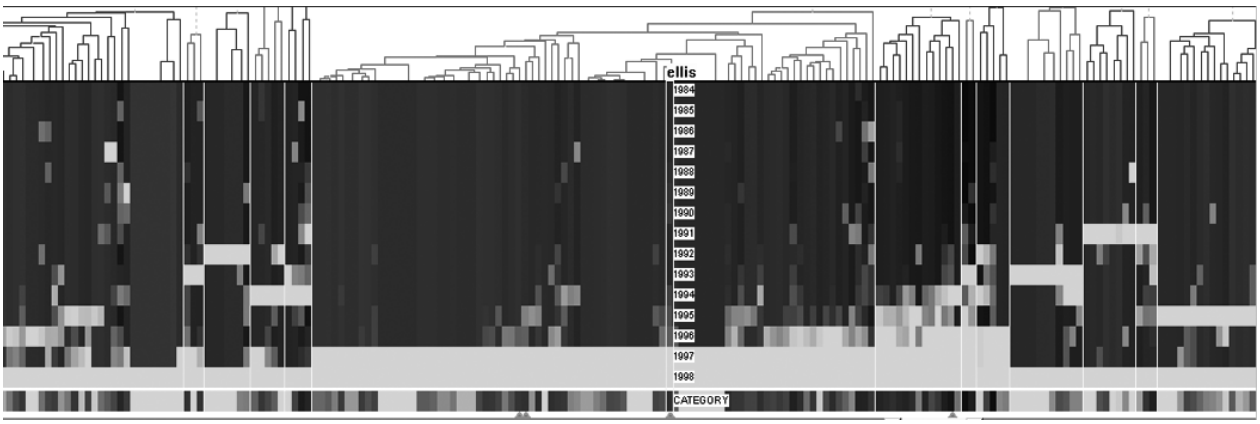

FIG. 8. A zoomed-in view of the dendrogram. The four relationships related to Shneiderman's interest in policy are denoted with triangles at the bottom of the graphic. One of these relationships (Ellis) is highlighted.

DOI: 10.1002/asi

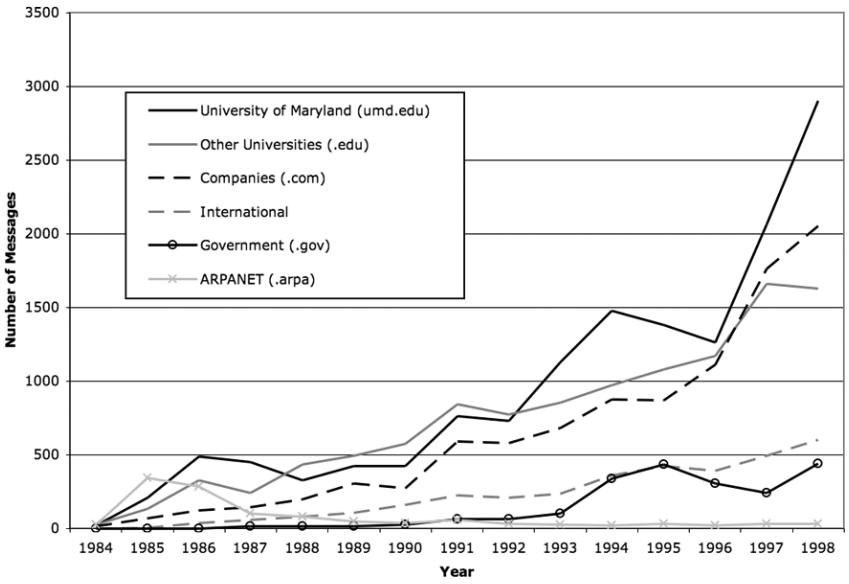

FIG. 9. Aggregate rhythms generated from domain names. The early ARPANET activity disappeared as it gave way to the newer domains. Most domain name activity increased over the years, but changes in the .gov activity are related to projects with government agencies.

however, a weakness of this approach is that not all of these clusters have meaning. For example, the algorithm finds three relationships that have peaks in the disparate years of 1988 and 1994. After exploring deeper into the e-mail content, it appears that is about all these relationships have in common.

Aggregating Related Rhythms

In addition to looking at the pattern of individual relationships, it also is a useful exercise to visualize rhythms of related aggregate relationships to see trends based on other attributes such as organization and location. For this corpus, we generate the aggregates from information contained within the e-mail headers. For each relationship, the most frequent e-mail address will represent that relationship's attributes. Of course, when dealing with an individual's e-mail archive, all of the addresses used by the owner should be disregarded. For each relationship, we extract organization names (IBM from user@*ibm*.com), organization type (educational from user@umd.*edu* versus commercial from user@spotfire.*com*) and country codes if present (Israel from user@technion.ac.*il*). With this extracted information, we illustrate some of the types of analysis that can be performed.

Although the number of active relationships increases over time, it became clear that many of Shneiderman's e-mail messages were still dedicated to relationships within his organization. Over the 15-year period, 24% of his e-mail was in communication with relationships at his own university, the University of Maryland. This percentage is comparable to the total fraction of messages in relationships with colleagues at other academics institutions (25%) and all corporations (23%), and double the number of messages beyond the U.S. borders (12%). Figure 9 shows a plot of the number of messages with each type of organization over the 15-year time period.

Figure 9 also shows how the contact base of international contacts grew over the 15-year time period. As Shneiderman's total number of messages grew, so did his correspondence with international contacts. Segmenting the data by country allows us to easily find the most popular international relationships. The top five countries are the United Kingdom (84 relationships), Canada (63), Germany (39), Israel (35), and Japan (31).

Grouping relationships by country allows explorers to notice trends present in Shneiderman's international rhythms. Countries such as Germany, Canada, Japan, and the United Kingdom have stable rhythms throughout most of the archive; however, other countries such as Australia, France, and Italy only grow toward the end of the archive. Other distinct profiles, like those of Austria and Finland, peak in intensity toward the middle of the archive and then fade as time goes on.

This approach allows explorers to find patterns and trends based on relationships sharing similar attributes; however, the e-mail address might not be an accurate representation of the relationship, thereby skewing the rhythms. Furthermore, individuals may change their organization and location over time, but our method will assign the relationship only its most frequent attributes over the duration of the archive.

Collaboration Rhythms

One important feature of e-mail is its ease of simultaneously distributing messages to more than one person. This is a typical activity when collaborating with colleagues, and these collaborations are evidenced by e-mail headers addressed to multiple people. To gain insights, we construct collaboration rhythms: rhythms characterized by the intensity

DOI: 10.1002/asi

of correspondence between two individuals, besides the archive owner, over time. Collaboration rhythms can be constructed by calculating the number of times two unique people are a part of the same conversation over the duration of the archive. These rhythms can be generated with an $O(N^2)$ algorithm which iterates through every e-mail address in the corpus that does not belong to the archive owner, and counts the number of times it is a part of an e-mail (e.g., listed on the to/from/cc lines of the e-mail header) with every other e-mail address in the corpus.

When plotting the collaboration rhythms of Shneiderman's archive, some interesting trends become evident. Most collaborations seemed to last less than 1 year, and it was rare for a collaboration to last more than 2 years. The collaboration rhythms with the most interesting patterns generally turned out to be mailing lists (e.g., a common poster to a particular list), as mailing lists have unique e-mail addresses as well. However, even with these shortcomings, it was easy to discern the top collaborators by glancing at the sharp peaks after superimposing all collaboration rhythms into one plot. These collaborations reinforce the notion that Shneiderman's intense e-mail relationships focus on coordination of distinct projects over time. Without collaboration rhythms, it would be hard to get a sense of the nature of collaborations between individuals in the archive.

A limitation of this approach is that if users change their e-mail addresses over time, the rhythms will be incomplete; however, folder metadata and the referencing user's full name from the e-mail header could help reduce the noise by creating more robust identities of users.

Rhythms of Relationships in Other Archives

A reasonable concern about our research methods is whether they are applicable to e-mail archives beyond the special case of the Shneiderman archive. Our attempts to replicate the analysis were limited by the absence of publicly available e-mail archives, which is partially attributable to the fact that copyright permissions would have to be obtained from every e-mail author (not just the recipient). Fortunately, the recently released e-mail archives of the Enron Corporation gave us a unique opportunity to apply our methods to a wholly different corpus.

Kenneth Lay was the Chairman and CEO of the Enron Corporation, and a snapshot of his e-mail, along with 150 other Enron employees, was released to the public (Cohen, 2005). In this archive, there were 18,945 messages in Lay's e-mail store, but over 73% of the messages were duplicates. This left us with 5,070 unique messages. The mail's dates ranged from March 1997 through December 2001; however, over 95% of the mail was from 2000 and 2001. Since the bulk of the e-mail archive spans only 2 years, rhythms based on yearly totals did not seem interesting. For this archive, we constructed rhythms based on weekly values for all 5 years of the archive. Mr. Lay did not file e-mail in folders based on relationships, so we assumed that each unique e-mail address also was a unique relationship, and thus a unique rhythm was constructed. Figure 10 shows 3,816 rhythms superimposed to illustrate the range of variation in the archive. Several of the most dominant rhythms are Steven Kean (Vice-President of Public Affairs), Richard Shapiro (Vice-President of Regulatory Affairs), and James Steffes (Vice-President of Government Affairs).

Of these rhythms, 2,935 (77%) were based on e-mail addresses from within the Enron Corporation (enron.com). A sample of other organizations that had multiple relationships with Lay were an international accounting firm (arthuranderson.com), a Texas-based law firm (velaw.com), a worldwide consulting firm (mckinsey.com), and a private all-girls Catholic school in Houston (duchesne.org). Although a

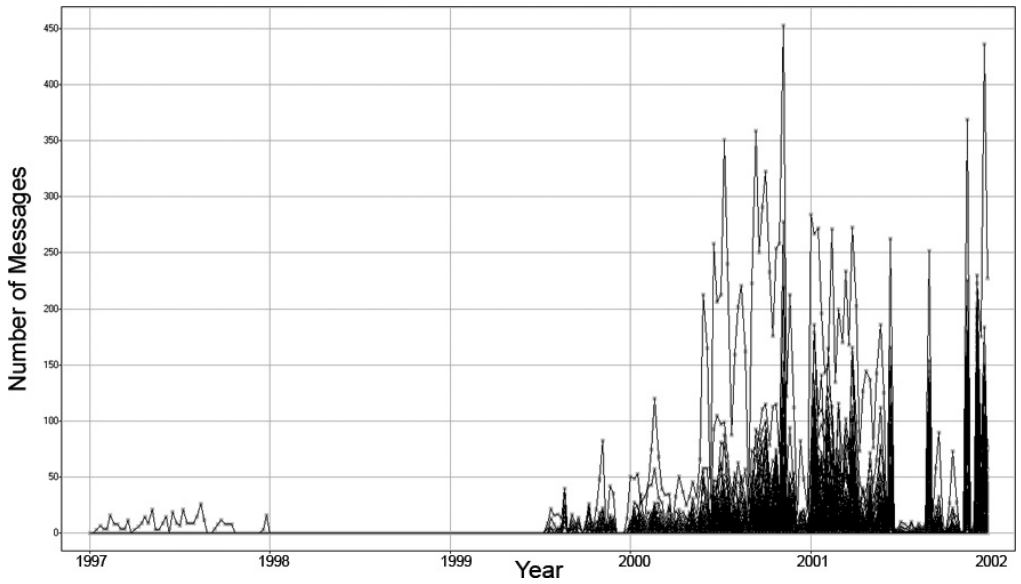

FIG. 10. Kenneth Lay's 3,816 rhythms of relationships superimposed, aggregated weekly from 1997 to 2001. The rhythms highlight the absence of e-mail in 1998 and early 1999. The most dominant rhythm in 2000 and 2001 was Stephen Kean, Enron's Vice-President of Public Affairs.

DOI: 10.1002/asi

rigorous analysis was not performed on the Lay archive, this example illustrates how our methods are applicable to corpora besides the Shneiderman archive.

Future Work

Rhythms of relationships offer a class of information that is hard to discern from keyword searching or reading the body of the e-mail messages; however, our rhythms will answer only a subset of questions that searchers may have. Our research interests are to build on the knowledge gained in this article and devise additional ways that searchers can learn more about the archive.

One weakness of our use of the clustering algorithms is that they do not cluster independent of time. For instance, if two relationships have identical curves over a time segment but occur in disparate years (e.g., one rhythm segment centers on 1989 vs. a second rhythm's center of 1996), our algorithms do not consider them similar. Interesting results can emerge by finding similar peaks and growths, such as determining if there is a typical rhythm associated with classes of people over time (e.g., a typical graduate-student curve) or if a certain initial pattern of activity predicts a durable or intense relationship.

The Shneiderman rhythms discussed in this article use a granularity of 1 year, which was motivated by our interest in understanding long-term rhythms; however, we suspect different evidence will emerge if the analysis were repeated with a granularity of months, weeks, or days. In the case of Shneiderman, we predict distinct trends of rhythms surrounding academic semesters, conferences, and weekends.

Although we believe our techniques are universal, so far they have only been rigorously tested on the Shneiderman e-mail archive. Our initial tests on the Enron archive show promise that similar success is achievable on archives of various durations and sizes.

Conclusion

Historians and social scientists believe that e-mail archives are important artifacts for understanding the individuals and communities they represent; however, there are currently few methods or tools to effectively explore these archives. This article presents a novel approach by analyzing the temporal rhythms of relationships in an e-mail archive. By visualizing these rhythms, important relationships become evident, searchers can find patterns of interest, and aggregate trends can be identified. We apply these techniques to the Shneiderman archive, and discover insights that may have been otherwise hidden.

Rhythms of relationships are an innovative way to understand e-mail archives; however, the novel approach also comes without rigorous testing. More evaluation is necessary, but the insights observed from the Shneiderman archive offer promising expectations. We feel the techniques we introduce help provide context that is necessary for historians and social scientists to make effective use of the archives. The number and size of e-mail archives will undoubtedly grow in future years, and searching them will become a more customary task. By presenting new ways to approach the exploration of e-mail archives, we provide a new step for effective exploration and also raise awareness for the difficult task of understanding e-mail archives.

Acknowledgments

We thank Susan Davis, Danyel Fisher, Mara Hemminger, Dave Levin, and Anthony Ramirez for their thoughtful comments on prior versions of the article. We also thank Jinwook Seo for providing assistance with Hierarchical Clustering Explorer. Adam Perer and Douglas Oard have been supported by DARPA cooperative agreement N66001002810 and the Joint Institute for Knowledge Discovery at the University of Maryland.

References

Baron, J.R. (1999). Email metadata in a post-Armstrong world. Proceedings of the 3rd IEEE Metadata Conference. Retrieved May 5, 2005, from http://www.computer.org/proceedings/meta/1999/papers/83/jbaron.html

Cohen, W. (2005). Enron email dataset. Retrieved May 5, 2005, from http://www.cs.cmu.edu/~enron/

Donath, J. (2004). Visualizing email archives [draft]. Retrieved May 5, 2005, from http://smg.media.mit.edu/papers/Donath/EmailArchives.draft.pdf

Ducheneaut, N., & Bellotti, V. (2001). Email as habitat: An exploration of embedded personal information management. Interactions, 8(5), 30–38.

Grieve, T. (2003). The decline and fall of the Enron empire. Salon.com, October 14, 2003. Retrieved May 5, 2005, from http://www.salon.com/news/feature/2003/10/14/enron/index_np.html

Johnson, S.C. (1967). Hierarchical clustering schemes. Psychometrika, 2, 241–254.

Kerr, B. (2003). Thread arcs: An email thread visualization. Proceedings of the 2003 IEEE Symposium on Information Visualization (p. 27). Los Alamitos, CA: IEEE Computer Society Press.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. Proceedings of the 8th international conference of ACM SIGKDD on Knowledge Discovery and Data Mining. New York: ACM Press.

Leuski, A., Oard, D.W., & Bhagat, R. (2003). eArchivarius: Accessing collections of electronic mail. Proceedings of the 26th annual ACM SIGIR Conference (p. 468). New York: ACM Press.

Li, W., Hershkop, S., & Stolfo, S.J. (2004). Email archive analysis through graphical visualization. Proceedings of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security (pp. 128–132). New York: ACM Press.

Mackay, W. (1988). More than just a communication system: Diversity in the use of electronic mail. Proceedings of the 1998 ACM conference on Computer-Supported Cooperative Work (pp. 344–353).

MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability (pp. 281–297). Berkeley, CA: University of California Press.

Rohall, S.L., Gruen, D., Moody, P., Wattenberg, M., Stern, M., Kerr, B., Stachel, B., Dave, K., Armes, R., & Wilcox, E. (2003). ReMail: A reinvented email prototype. Proceedings of ACM Human Factors in Computing Systems (pp. 791–792). New York: ACM Press.

Sack, W. (2000). Discourse diagrams: Interface design for very large scale conversations. Proceedings of the 33rd Hawaii International Conference on System Sciences (p. 3034). Los Alamitos, CA: IEEE Computer Society Press.

Seo, J., & Shneiderman, B. (2002). Interactively exploring hierarchical clustering results. IEEE Computer, 35(7), 80–86.

DOI: 10.1002/asi

Smith, M. (1999). Invisible crowds in cyberspace: Measuring and mapping the social structure of USENET. In M. Smith and P. Kollock (Eds.), Communities in cyberspace (pp. 195–219). London: Routledge Press.

Smith, M. (2002). Tools for navigating large social cyberspaces. Communications of the ACM, 45(4), 51–55.

Spotfire. (2005). Spotfire DecisionSite. http://www.spotfire.com.

Tyler, J.R., & Tang, J.C. (2003). When can I expect an email response? A study of rhythms in email usage. Proceedings of the European Conference on Computer-Supported Cooperative Work 2003 (pp. 239–258). Dordrecht, The Netherlands: Kluwer Academic.

Tyler, J.R., Wilkinson, D.M., & Huberman, B.A. (2003). Email as spectroscopy: Automated discovery of community structure within organizations. Proceedings of Communities and Technologies. Dordrecht, The Netherlands: Kluwer Academic.

Venolia, G., & Neustaedter, C. (2003). Understanding sequence and reply relationships within email conversations: A mixed-model visualization. Proceedings of ACM Human Factors in Computing Systems (pp. 361–368). New York: ACM Press.

Viegas, F., Boyd, D., Nguyen, D., Potter, J., & Donath, J. (2004). Digital artifacts for remembering and storytelling: PostHistory and social network fragments. Proceedings of the 37th Hawaii International Conference on System Sciences (p. 40109a). Los Alamitos, CA: IEEE Computer Society Press.

Wattenberg, M. (2001). Sketching a graph to query a time series database. Proceedings of ACM Human Factors in Computing Systems (pp. 379–380). New York: ACM Press.

Whittaker, S., & Sidner, C. (1996). Email overload: Exploring personal information management of email. Proceedings of ACM Human Factors in Computing Systems (pp. 276–283). New York: ACM Press.

DOI: 10.1002/asi