Making Sense of Archived Email: Exploring the Enron Collection with NetLens

Hyunmo Kang,[a] Catherine Plaisant,[a] Tamer Elsayed,[b] and Douglas W. Oard[a,c]
[a]UMIACS Human-Computer Interaction Laboratory,
[b]Computer Science Department, and [c]College of Information Studies
University of Maryland, College Park, MD 20742
{kang, plaisant, telsayed, oard}@cs.umd.edu

Abstract

Informal communications media pose new challenges for information systems design, but the nature of informal interaction offers new opportunities as well. This paper describes NetLens-Email, a system designed to support exploration of the content-actor network in large email collection. Unique features of NetLens-Email include close coupling of orientation, specification, restriction and expansion and introduction and incorporation of a novel capability for iterative projection between content and actor networks within the same collection. Scenarios are presented to illustrate the intended employment of NetLens-Email, and design walkthroughs with two domain experts provide an initial basis for assessment of the suitability of the design by scholars and analysts.

1. Introduction

Systems designed to support information seeking processes have traditionally focused on documents such as books, scholarly articles and news stories that are written in a formal style and carefully edited. The emergence of the Web has brought new challenges (e.g., limiting adversarial manipulation of search results from "search engine optimization"), new opportunities (e.g., from the rich network of interconnections between Web pages), and increased emphasis on certain types of information needs (e.g., "navigational queries" seeking a service rather than a document).

We now find ourselves on the verge of another such transformation as increasing quantities of informally exchanged messages are created, retained, and ultimately made searchable. Hardly a week goes by without mention in the press of the use of stored email messages as evidence in some investigation. Retention of instant messaging records is now required in certain regulatory settings (e.g., for securities dealers). Rapid and affordable document imaging services have facilitated mass digitization of memoranda, notes, and other informal messages that previously were available only in hardcopy form when their release is sought incident to litigation or a regulatory investigation. This transformation is not limited to written sources: advances in automatic speech recognition have already rendered some informal speech transcribable with sufficient accuracy to support some information access tasks (e.g., query-based search or automatic topic classification). Now is therefore a propitious time to explore the design of systems to support access to these types of retained records from informal interaction.

Informally written content has long been a focus of work in an area broadly known as "personal information management," the distinguishing characteristic of which is that the user participated in the recorded interactions. Our interest, by contrast, is in "archival access" applications in which the user was a non-participant. Representative users in our context would include historians and other scholars seeking to make sense of past events, archivists and records managers who are responsible for managing retention and description of specific collections, lawyers tasked with disclosing relevant evidence in response to a plaintiff's "discovery" request, and law enforcement or intelligence analysts seeking to understand specific intentions, actions, or events. Two important characteristics of those situations are that (1) the amount of informal interaction is often larger than can be understood by any individual, and (2) the people seeking to make sense of what they find may initially be unfamiliar with some of the people mentioned in the collection and some of the topics discussed there.

This formulation leads us naturally to think of our user's task as "sense-making" rather than simply as search. By sense-making in this context we mean a directed process in which the user assembles evidence to construct a personal understanding of specific aspects of the meaning of recorded interactions (Dervin and Foerman-Wernet. 2003). In essence, sense-making is better thought of as learning than as simply finding. For example, when viewing an email inquiring

about when someone plans to arrive, the user might then want to know what the purpose of the visit had been. The process of sense-making is iterative, evolving through progressively more focused information needs that themselves may be recursively decomposed into smaller and even more tightly focused information needs. Multiple sources of information are typically consulted, although at this point in our work we have chosen to focus on exploration of a single collection.

An important feature of informal interaction is the complementary relationship between identity and topicality. For example, if we want to find White House emails that address tobacco policy, it would help to know which staff members had policy responsibilities. Similarly, if we want to know how someone spent their weekend, that information is more likely to be found in a message sent to a friend than in a message broadcast to a departmental mailing list. Focusing on identity is not new, of course. For example, source authority has long been recognized as an important factor in information seeking behavior (e.g., (Meho and Tibbo, 2003)). Informal interaction brings an explosive increase in the number of sources, however. As a result, making sense of identity can be just as challenging, and just as important, as making sense of topical content. For example, the CMU version[1] of the Enron email collection (Klimt and Yang, 2004) contains a snapshot of email retained by just 149 individuals, but together those half-million emails contain more than 123,000 unique email addresses.[2] Of course, any specific sense-making task will focus on the roles, relationships, interests and activities of just a few of those people. Search and sense-making are therefore close cousins.

If we view sense-making as a process engaged in by people with the help of machines, then supporting effective interaction naturally becomes a focal issue when designing the machine. In this article, we present an exploration support system, which we call *NetLens*, that employs a factored design in which the same basic functions can be applied to a variety of sense-making problems for which identity and content can productively be reasoned about jointly (e.g., studying the growth and decline of online communities, or fostering knowledge management within organizations). NetLens integrates five key capabilities: orientation, specification, restriction, expansion, and projection; we describe each of these in Section 4. The system was initially applied

[1] http://www.cs.cmu.edu/~enron

[2] As points of comparison with topical sense-making, both the Bible and the complete works of Shakespeare each contain fewer than 25,000 unique words.

to a content-author network constructed from scholarly publications from the ACM digital library (Kang et al., 2007). We focus here on the new capabilities that we added to create a "NetLens-Email" system, including multi-scale (message and thread) content representation, more comprehensive representations of identity, and (for a subset of the collection) rich annotation. Our principal interest at this stage in our research is eliciting insight into the suitability for this task of the NetLens interaction design. To gain that insight, we conducted design walkthroughs with two users who brought expertise and experience in some specific types of sense-making tasks that we seek to support. In those walkthroughs, the users explored the Enron collection.

The remainder of this paper is organized as follows. Section 2 reviews prior work on email analytics and visualization. Section 3 introduces the Enron collection and describes the preprocessing that we performed on that collection. Section 4 describes the NetLens-Email system in detail. Section 5 then contrasts NetLens-Email with the design of an interactive ranked retrieval system that we use as a baseline against which the more advanced capabilities of NetLens-Email are contrasted. Section 6 presents results from design walkthroughs with two experts, a professional archivist and an experienced analyst, each of whom first used our baseline system and then NetLens-Email. We conclude in Section 7 with remarks on the implications of this work for future research on helping non-participants make sense of large collections of recorded informal interaction.

2. Related Work

Our NetLens-Email system integrates analytic and visualization functions, so we review prior work on those topics in the context of email.

2.1 Email Analytics

A key task that has been extensively investigated is discovery of relationships between the individuals who participated in construction of an email collection; this is a form of social network analysis. For example, McArthur and Bruza (2003) found explicit and implicit connections between people by mining semantic associations inferred from their email communications. This kind of link characterization can in turn serve as a basis for more sophisticated inference. For

example, Leuski (2004) analyzed the patterns of speech acts in incoming and outgoing emails to infer the roles of specific individuals from a small collection of 500 emails. Tyler et al. (2005) undertook a similar task on a far larger collection, using centrality statistics to identify communities and leadership roles from a graph representing information flows evident in a collection of nearly one million email headers. There has also been some work on associating email addresses with mentions of individuals in the body text of email messages that offers the potential to extend social network analysis to include mentions in addition to interactions (Elsayed et al., 2008; Diehl et al., 2006; Minkov et al., 2006).

Information flows within a social network have also been used in conjunction with topic characterization. For example, McCallum et al. (2007) developed a Bayesian network that learned a topic distribution for communication between two entities based on the content of the messages sent between them. One important application for the results of that type of analysis is finding experts on a specific topic (Craswell et al., 2005). Two basic approaches to exploiting topic matching for that problem have been tried. The first builds a profile for each individual from emails they have sent and received, then searches those profiles based on a query. The alternative is to index each email separately, identify emails that match the query, and then to infer the most likely experts based on those highly ranked emails. Balog et al. (2006) systematically compared the two approaches using a large collection of emails sent to World Wide Web Consortium mailing lists, finding the second approach to be both more effective and more efficient. Zhang and Ackermann (2005) adopted a complementary perspective, investigating expert finding based on social network analysis. In experiments with the Enron email collection they found that asking well connected neighbors (i.e., those who sent messages to many different correspondents) to help identify an expert on some specific topic would in be a good strategy (e.g., that it would generally be better than asking people whose expertise is similar to yours). Evidence from email has also been used in conjunction with other sources. For example, Culotta et al. (2004) used the Web to find contact information for people whose names and email addresses were extracted from email headers.

A problem known as “e-discovery” in the context of civil litigation has emerged as a driving application for effective and efficient search in large email collections (and many other types of

digital data). As digital records displace physical records of individual, corporate and government activity, massive collections are being amassed that exceed what the manual review techniques of an earlier era can accommodate (Baron and Thompson., 2007). Development of search technology tailored to the requirements of this task is being investigated by participants in the Text Retrieval Conference' legal track (Oard et al., 2008).

Although email addresses can provide a strong basis for modeling identity, intentional deception can be a problem in some applications. Some early studies on authorship attribution in email were conducted by de Val (de Val, 2000; de Val and Anderson, 2001). In their first study, a support vector machine was used with structural, lexical and stylistic features to distinguish between five authors in a collection of 247 emails on a single topic. Their subsequent study extended the content-independent features to build classifiers that could discriminate between authors in the presence of three different topics. More recently, Zheng et al. (2006) developed a framework for authorship identification using lexical, syntactic, structural and stylistic features with inductive learning algorithms to identify authorship of Usenet messages, which resemble mailing list postings. The approach was able to identify authors with accuracies between 70% and 95%. Deception detection in email content has also received recent attention. For example, Keila and Skillicorn (2005) applied a model based on patterns of word usage to detect deceptive emails in Enron collection.

2.2 Email Visualization

Visualization has been used to analyze both personal email collections and public archives of threaded discussions that resemble the ways mailing lists are often used. The focus of research with personal email collections has most often been personal information management. For example, Remail provided overviews of correspondents and messages to help spot those with similar attributes (Rohall et al., 2003). PostHistory was a personal information management tool for using timelines and contact overviews to help generate insights that would be socially relevant to the owner of an email collection (Viegas et al., 2004). Social Network Fragments complemented that approach, identifying communication clusters in social networks of authors (Viegas et al., 2004). There has also been work on showing the structure of threaded email conversations over time (Kerr, 2003; Venolia and Neustaedter, 2003). Email visualization has

recently started to receive some attention from researchers seeking to support retrospective analysis as well. For example, Perer et al. used timelines to study the rhythms in email use over time with a case study based on 20 years of a single individual's email (Perer and Shneiderman, 2005; Perer et al, 2006). Perer and Smith looked at three simple visualization of emails to construct portraits of email practices and collected feedback from 8 users looking at their own email store (Perer and Smith, 2006).

Until recently, data availability issues made it difficult to study archives of emails from groups of people. The eArchivarius project showed clusters based on content or co-addressing, along with timelines and biographies to explore a small set of government email that had been released in response to Freedom of Information Act requests (Leuski et al., 2003). For the half-million-document Enron collection, an interactive graph exploration tool was developed by Heer (2005) to explore the social network of correspondents. To obtain even larger collections, researchers have most often turned to Usenet and to public mailing lists. Those services are often used somewhat differently from personal email, but the common message format at least makes it possible to apply similar tools. On a larger scale, Sack's (2000) "Conversation Map" reveals the long-term structure of Usenet discussion groups by using social network diagrams, lists of discussion themes, and a semantic network, and other researchers have applied Treemaps in the Netscan project to visualize Usenet postings (Smith and Fiore, 2001).

Closely coupling analysis capabilities with visualization, an approach often referred to as "visual analytics," (Thomas and Cook, 2005) can yield capabilities that are greater than the sum of the parts. Of course, all analysis results are displayed in some way, and all visualization is based on some analysis. So the moniker "visual analytics" is best used to describe collaborations that extend beyond the traditional boundaries of those individual fields. Early examples include the use of dimensionality reduction to visualize thematic relationships in vector spaces (Wise et al., 1995) while newer work combines multiple approaches (e.g., (Proulx et al., 2006, Seo and Shneiderman, 2005)).

3. Preprocessing

The capabilities of our NetLens-Email system are grounded in extensive preprocessing for representation of content and identity, both in a form designed to support rapid automated response during interactive use and in a form designed for on-demand display to the user. This section first describes how we represent content, and then how we represent identity.

3.1. Representing Content

We started with the CMU version of the Enron collection, which contains 517,431 email messages, without attachments (Klimt and Yang, 2004). We augmented this email content with 93 transcribed recordings of telephone conversations between Enron energy traders and others, rendered in a format similar to that used for email (Goldstein et al., 2006). In preprocessing, we automatically detected salutation and signature lines and separated new body-text (i.e., text actually written by the sender of the email) from the quoted text (i.e., sections of a forwarded or replied-to message that are automatically copied into the message by the sender's email client). Duplicate messages were removed. We considered two emails to be duplicates if they have exactly the same: (1) email addresses of sender and receivers, (2) subject, and (3) body (after being tokenized and stemmed). This process resulted in detection of 268,980 duplicate emails; about 52% of the collection (because many messages were stored both by their sender and by some recipients). Subsequent processing was restricted to the remaining 248,451 unique emails. We recorded the folder in which each message had been found as additional header metadata; because of duplicate detection this field could contain more than one folder.

Although thread relationships are sometimes encoded by email software in an optional in-reply-to header, that is not the case in the CMU version of Enron collection. We therefore adopted a technique similar to that introduced by Lewis and Knowles, in which quoted headers and quoted text are used to form a retrieval query (using Lucene[3], an open source Java library for text retrieval applications) to identify the most likely parent email in the same thread (Lewis and Knowles, 1997). When quoted text was not available, we used the subject line to recognize thread relationships (by removing terms such as “re:” and “fwd:” and then looking for identical nontrivial subject lines). We defined email threads as tree structures that include all emails that could be

[3] Available at http://lucene.appache.org

linked by parent-child relationships detected in these ways. Unlike typical (reply chain) threads, our threads also include forwarding relationships. The resulting threads sometimes contain parts of more than one actual conversation; more robust reconstruction techniques are, of course, possible (e.g., Yeh, 2006).

We built separate indexes for messages and for threads. Any text found in the subject field, in the main body, or in quoted text, was stemmed, English stop-words were removed, and the resulting terms were indexed using Lucene. Lucene implements a variant of the vector space model, with special features for handling multi-field documents. For all messages, we indexed the subject line, text quoted from other messages, other (new) body text, date, time, sender, recipients (to:, cc:, and bcc:), message type (email or phone call), and mentioned persons. To automatically identify persons who were mentioned in the header, body or quoted text or each message, we used LingPipe,[4] a named-entity extraction tool.

Two parts of the collection had additional annotations available that we also indexed. In the 93 transcribed telephone conversations, names of mentioned persons and some important words (e.g., "email", "Enron", and "Ricochet") were manually annotated (Goldstein et al., 2006). For 1,700 of the messages, additional manually annotated faceted descriptive metadata was available from UC Berkeley.[5] This metadata includes coarse genre (company business, purely personal, etc.), types of attachments[6] (press release, legal document, etc.), primary topics keyed to a collection-specific ontology (California energy crisis, legal advice, trip report, etc.), and emotional tone (secrecy, worry, gratitude, etc.).

All of these representations were available to support both automated processing and display to the user. In addition, we automatically constructed one representation specifically for display to the user: a 100-word generic (i.e., not query-specific) summary of each thread (Zajic et al., 2008). The constituent emails in the thread were first pre-processed to remove quoted text and to identify and syntactically parse each sentence. The Hedge Trimmer system, which had been originally designed for summarization of news and later adapted to email summarization, was then used to

[4] http://alias-i.com/lingpipe/
[5] http://bailando.sims.berkeley.edu/enron_email.html
[6] The attachments themselves are not included with the CMU collection.

generate multiple compressed candidates for each sentence. A sentence selector based on maximum marginal relevance was used to select among the resulting compressed candidates for inclusion in the summary. The goal of this process was to maximize inclusion of sentences describing central concepts from the thread, while minimizing redundancy within the summary.

3.2. Representing Identity

The CMU version of the Enron collection contains 133,581 unique email addresses. We collapsed this set to 123,783 identities by automatically recognizing co-referent addresses using header co-occurrence and exact full-name matching. As identity attributes, we used domain names extracted from email address (enron.com, (other).com, .edu, .gov, etc.), person names, and job titles, statistically improbable phrases, number of correspondence partners, connectance, and centrality. Person names were automatically associated with email addresses based on co-occurrence in email headers, or in automatically detected salutation or signature lines (Elsayed and Oard, 2006). Job titles were manually annotated only in the 1,700-message UC Berkeley subset of the email collections based on database content that had been released as a part of the Enron investigation (Hemminger, 2005). The number of correspondence partners was computed as the number of unique identities to which they had sent at least one message or from which they had received at least one message in the collection. Connectence (a clustering coefficient) was computed as the proportion of the person's correspondence partners (computed as above) who had also corresponded directly with one another. Centrality was computed as the average of the shortest path lengths from the identity to every other identity in the communications graph (where any message between two identities defined an undirected link). Collectively, we refer to the number of correspondence partners, connectence, and centrality as "social network statistics."

As with content representation, we built one identity representation specifically for display to the user. We call this representation a "bio" (short for biography). The bio was designed as an easily skimmed summary of useful information about a person, including their full name, job title (when known), email address, commonly used signature lines, and most frequent correspondence partners (shown as email addresses), the social network statistics, and statistically improbable

phrases (word sequences that were much more commonly used by one identity than by any other).[7]

4. The NetLens-Email System

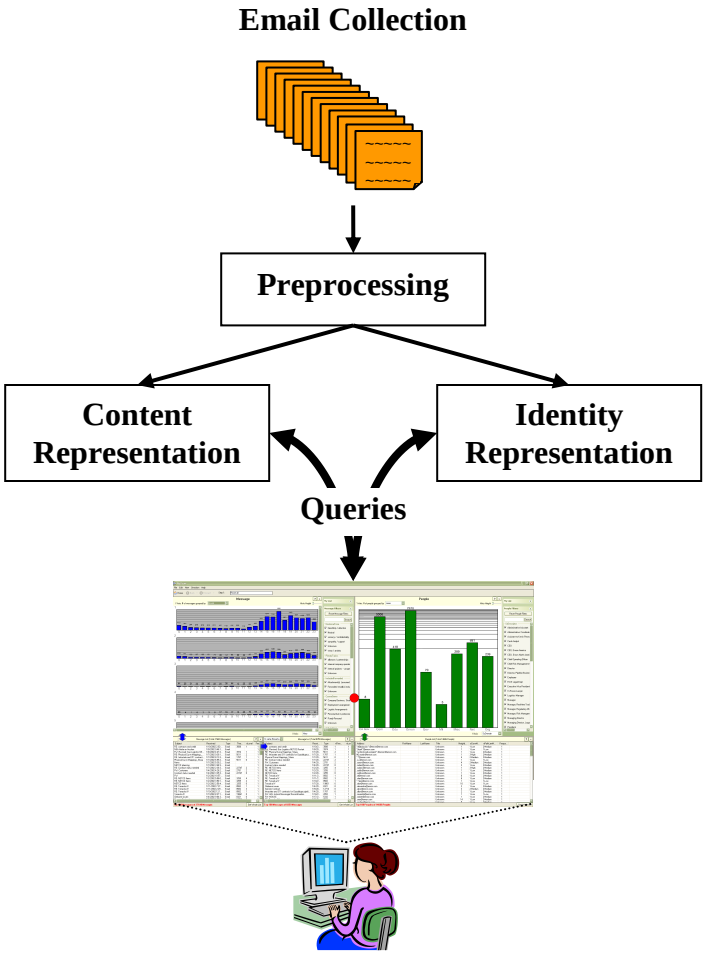


Figure 1. Design perspective of NetLens-Email

NetLens was designed to serve as a research prototype that allows us to try out different interface designs, interaction strategies, and component capabilities for applications in which identity and content can both serve important access points. NetLens-Email is an evolutionary development of NetLens (Kang et al., 2007), itself based on PaperLens (Lee et al., 2005) both originally designed to help users explore portions of the ACM Digital Library. Figure 1 illustrates the design perspective of NetLens-Email. NetLens employs a factored design in which the data store is

[7] http://code.google.com/p/statistically-improbable-phrases/

coupled to the user interface through a service architecture. A relational database serves the primary data store for items that can be computed in advance, with additional capabilities (e.g., expansion based on content similarity) implemented as separate services that are invoked on demand.

4.1 NetLens Capabilities

NetLens provides users with five fundamental capabilities:

Orientation: At first, users see an overview of the email collection both in terms of content and identity. They can see a visual representation of the distribution of the data over a variety of attributes (e.g., distribution of the number of emails by year, or by the time of the day), or distribution of people by status (e.g., sender vs. receiver, or inside vs. outside of Enron).

Specification: Users often start by specifying something about what they are looking for, either in the content or among the people sending and receiving (or mentioned in) the email. This might be a fairly specific reference (e.g., the full name of a person, or a specific phrase that might be found in some email), or the starting point may be more general (e.g., the name of a company where some people work, or a relatively common word that might have been used in several contexts). Specification produces a result set, optionally ranked by a user-specified criterion (e.g., a list of people working for a company, optionally ranked by their centrality in the social network among that company's employees) Details on the "rank-by-feature framework" that motivates this design can be found in (Seo and Shneiderman, 2005).

Restriction: Once a set has been created, it can be further restricted by specifying additional required characteristics to create a subset. Simple examples include truncating a ranked set at some fixed cutoff (e.g., deleting people who sent few emails), or specifying an additional required characteristic (e.g., limiting the time frame being considered).

Expansion: The size of a set can also be expanded to include related materials. This can be done in several ways (e.g., addition of people who frequently exchange email with members of the set,

or addition of email messages that use terms in similar ways). When expansion is performed, the degree of similarity becomes an additional criterion that can be used to rank the expanded set.

Projection: Some or all of the first four capabilities can be found in many information access applications (for example, as collection overviews, ranked retrieval, search-within result set, and relevance feedback). It is the addition of projection that distinguishes our work by leveraging the content-actor framework that lies at the core of NetLens. In brief, sets based on content can be projected to create corresponding sets based on identity, and vice versa. For example, the recipients of emails found in a content set define an identity set, and the emails sent by people in an identity set define content set. This capability is described in greater detail below.

The NetLens architecture consists of two key components: NetLens Model and NetLens Viewer. NetLens Model manages all the data-related processes; NetLens Viewer can therefore request needed data from NetLens Model without handling integration details for the specific database management system that NetLens Model uses. All of NetLens is written in C#. The open source Piccolo toolkit (Bederson et al., 2004) is used by NetLens Viewer to implement the two overview histograms. The present implementation of NetLens Model relies on the freely available MySQL database management system.

4.2 NetLens Interaction Design

Figure 2 shows the NetLens-Email user interface. The display is divided into two symmetric sections. On the left is the content (i.e., the email messages), and on the right are the actors (i.e., people, represented by their email addresses) associated with those emails. Each side includes four principal regions: (1) a query box at the top right, (2) an overview (a histogram to show the distribution of items over the selected attributes) at the top left, (3) filter controls to the right of the overview, and (4) one or more (optionally ranked) item lists below the overview. Within each side (content on the left, people on the right), the four regions are tightly coupled: actions in one region are immediately reflected in the other regions on that same side.

The pull-down menu at the top left in the overview region provides the user with control over the feature(s) used to group items on the X-axis of the overview histogram, thereby allowing selection

of views appropriate to the user's stage in their sense-making task. For example, in Figure 2, emails are grouped into 24 hour range based on their sent/received time for each month in 2002 to help identify unusual temporal patterns of email exchange in the data set. The overview serves both as a result display (using a histogram to show the distribution of items over the selected attributes) and as a restriction control (interpreting selection of one or more histogram bars as a filter to remove from the item list(s) all items that are not associated with selected bars). Users can also collect any entity instances (i.e. either email or person), save them to "my list", and then use this collection as a filter in the overview panel. Restrictions specified using the filter region, the overview histogram, or my list, are cumulative.

The sort order of the item list can be changed from the initial default (which varies by the type of list) by clicking on any field header once (for ascending order) or twice (for descending order), a commonly used process that we would expect many users to have experienced with other applications. Details for individual items (summaries or biographies, respectively) are shown on demand (in response to a double click in the item list) in a pop-up window that remains active until manually closed (thus supporting comparison of multiple items). Further drill-down from the summary to full text is also available on the content side if additional details are needed.

Initially the two sides are uncoupled; interaction with the content side would not affect the people side, and vice versa. This is indicated by the solid red circle at the center of the display in Figure 2. Users can, however, establish unidirectional projection (with a right-click on that circle). When projection is enabled, the circle is replaced by a solid arrow showing the projection direction and the user-selected basis for that projection operation (Senders of messages, receivers of messages, messages received by and messages sent by). A right-click in any other region (i.e., anywhere other than directly over the projection symbol) will instead activate a pop-up menu to control expansion operations.

A query history management system records every interaction users make in NetLens and automatically annotates each interaction to help users remember earlier interactions. This provides support both for confirming or correcting previous interactions and for documenting interaction strategies.

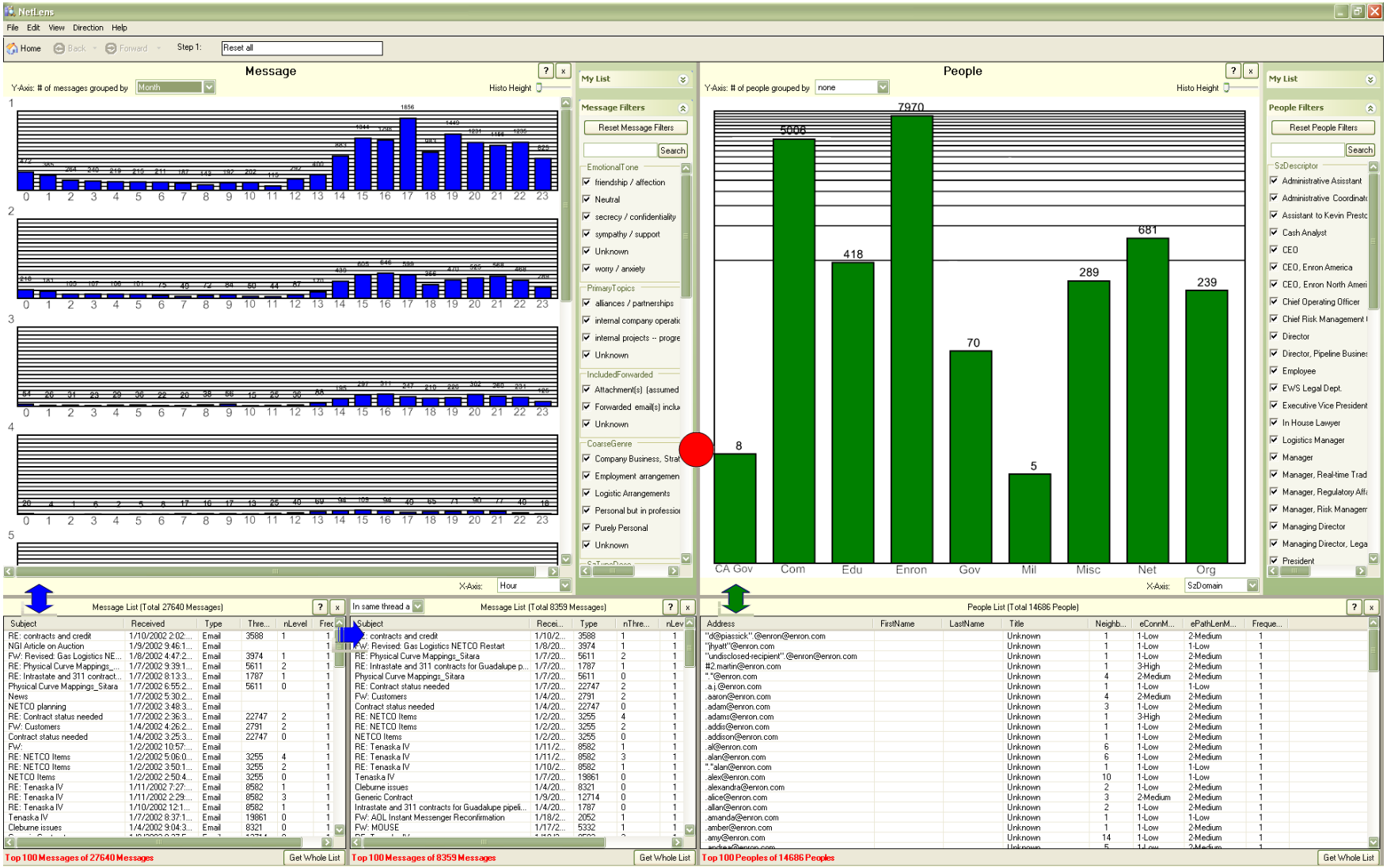

NetLens
File Edit View Direction Help
Home Back Forward Step 1: Reset all
Message
Y-Axis: # of messages grouped by Month
Histo Height
My List
Message Filters
Reset Message Filters
Search
EmotionalTone
friendship / affection
Neutral
secrecy / confidentiality
sympathy / support
Unknown
worry / anxiety
PrimaryTopics
alliances / partnerships
internal company operati
internal projects -- progre
Unknown
IncludedForwarded
Attachment(s) (assumed
Forwarded email(s) inclu
Unknown
CoarseGenre
Company Business, Strat
Employment arrangemen
Logistic Arrangements
Personal but in professio
Purely Personal
Unknown
X-Axis: Hour
People
Y-Axis: # of people grouped by none
Histo Height
My List
People Filters
Reset People Filters
Search
SzDescriptor
Administrative Assistant
Administrative Coordinat
Assistant to Kevin Presto
Cash Analyst
CEO
CEO, Enron America
CEO, Enron North Ameri
Chief Operating Officer
Chief Risk Management
Director
Director, Pipeline Busines
Employee
EWS Legal Dept.
Executive Vice President
In House Lawyer
Logistics Manager
Manager
Manager, Real-time Trad
Manager, Regulatory Affa
Manager, Risk Managem
Managing Director
Managing Director, Lega
President
8
5006
418
7970
70
5
289
681
239
CA Gov
Com
Edu
Enron
Gov
Mil
Misc
Net
Org
X-Axis: SzDomain
Message List (Total 27640 Messages)
Subject Received Type Thre... nLevel Freq
RE: contracts and credit 1/10/2002 2:02:... Email 3588 1 1
NGI Article on Auction 1/9/2002 9:46:1... Email 1
FW: Revised: Gas Logistics NE... 1/8/2002 4:47:2... Email 3974 1 1
RE: Physical Curve Mappings_... 1/7/2002 9:39:1... Email 5611 2 1
RE: Intrastate and 311 contract... 1/7/2002 8:13:3... Email 1787 1 1
Physical Curve Mappings_Sitara 1/7/2002 6:55:2... Email 5611 0 1
News 1/7/2002 5:30:2... Email 1
NETCO planning 1/7/2002 3:48:3... Email 1
RE: Contract status needed 1/7/2002 2:36:3... Email 22747 2 1
FW: Customers 1/4/2002 4:26:2... Email 2791 2 1
Contract status needed 1/4/2002 3:25:3... Email 22747 0 1
FW: 1/2/2002 10:57:... Email 1
RE: NETCO Items 1/2/2002 5:06:0... Email 3255 4 1
RE: NETCO Items 1/2/2002 3:50:1... Email 3255 2 1
NETCO Items 1/2/2002 2:50:4... Email 3255 0 1
RE: Tenaska IV 1/11/2002 7:27:... Email 8582 1 1
RE: Tenaska IV 1/11/2002 2:29:... Email 8582 3 1
RE: Tenaska IV 1/10/2002 12:1... Email 8582 1 1
Tenaska IV 1/7/2002 8:37:1... Email 19861 0 1
Cleburne issues 1/4/2002 9:04:3... Email 8321 0 1
Top 100 Messages of 27640 Messages
Get Whole List
In same thread a
Message List (Total 8359 Messages)
Subject Recei... Type nThre... nLev
RE: contracts and credit 1/10/2... 3588 1 1
FW: Revised: Gas Logistics NETCO Restart 1/8/20... 3974 1 1
RE: Physical Curve Mappings_Sitara 1/7/20... 5611 2 1
RE: Intrastate and 311 contracts for Guadalupe p... 1/7/20... 1787 1 1
Physical Curve Mappings_Sitara 1/7/20... 5611 0 1
RE: Contract status needed 1/7/20... 22747 2 1
FW: Customers 1/4/20... 2791 2 1
Contract status needed 1/4/20... 22747 0 1
RE: NETCO Items 1/2/20... 3255 4 1
RE: NETCO Items 1/2/20... 3255 2 1
NETCO Items 1/2/20... 3255 0 1
RE: Tenaska IV 1/11/2... 8582 1 1
RE: Tenaska IV 1/11/2... 8582 3 1
RE: Tenaska IV 1/10/2... 8582 1 1
Tenaska IV 1/7/20... 19861 0 1
Cleburne issues 1/4/20... 8321 0 1
Generic Contract 1/9/20... 12714 0 1
Intrastate and 311 contracts for Guadalupe pipeli... 1/4/20... 1787 0 1
FW: AOL Instant Messenger Reconfirmation 1/18/2... 2052 1 1
FW: MOUSE 1/17/2... 5332 1 1
Top 100 Messages of 8359 Messages
Get Whole List
People List (Total 14686 People)
Address FirstName LastName Title Neighb... eConnM... ePathLenM... Freque...
"d@piassick" @enron@enron.com Unknown 1 1-Low 2-Medium 1
"jhyatt" @enron.com Unknown 1 1-Low 1-Low 1
"undisclosed-recipient" @enron@enron.com Unknown 1 1-Low 2-Medium 1
#2.martin@enron.com Unknown 1 3-High 2-Medium 1
" "@enron.com Unknown 4 2-Medium 2-Medium 1
.a.j.@enron.com Unknown 1 1-Low 1-Low 1
.aaron@enron.com Unknown 4 2-Medium 2-Medium 1
.adam@enron.com Unknown 3 1-Low 2-Medium 1
.adams@enron.com Unknown 1 3-High 2-Medium 1
.addis@enron.com Unknown 1 1-Low 2-Medium 1
.addison@enron.com Unknown 1 1-Low 2-Medium 1
.al@enron.com Unknown 6 1-Low 2-Medium 1
.alan@enron.com Unknown 6 1-Low 2-Medium 1
."."alan@enron.com Unknown 1 1-Low 1-Low 1
.alex@enron.com Unknown 10 1-Low 1-Low 1
.alexandra@enron.com Unknown 2 1-Low 2-Medium 1
.alice@enron.com Unknown 3 2-Medium 2-Medium 1
.allan@enron.com Unknown 2 1-Low 2-Medium 1
.amanda@enron.com Unknown 1 1-Low 1-Low 1
.amber@enron.com Unknown 1 1-Low 2-Medium 1
.amy@enron.com Unknown 14 1-Low 2-Medium 1
Top 100 Peoples of 14686 Peoples
Get Whole List

Figure 2. (on previous page) NetLens-Email has two symmetric windows. The left is for Content (emails) and the right for Actors (people). Each side is further divided into panels; overview at the top, filters on the right, and lists at the bottom. Here the (left) Content side has two item lists, one for individual messages and one for threads. The single item list on the (right) Actor side shows email addresses for senders or recipients. The (left) Content overview panel shows the distribution of emails by time of day over 24 hours (with peaks near 5 PM), aggregated into one histogram for each months. On the (right) Actor side, the overview shows the distribution of top-level email domains (on a logarithmic scale).

4.3 Adapting NetLens to Email

To implement the services required for NetLens-Email, we started by building computational representations of content and identity as described in section 3.

Two types of services are provided in NetLens-Email:

Search: Identity searches are based on approximate string matching (e.g., Ralph and Ralf will match more closely than George and Harry) or on communications patterns (e.g., finding frequent correspondents). Content searches are based on term overlap and term importance measures (i.e., rare terms receive more weight) using standard information retrieval techniques. NetLens-Email supports two different kinds of content search; keyword search and similar-email search. With keyword search, users can retrieve a set of emails using a free-text query. With the similar-email search feature in NetLens, users can retrieve a set of emails similar to a given email (or a given set of emails). When a set of emails is specified, the new (not quoted) body text of each email is extracted and concatenated to construct a (rather long) text query that is then used to retrieve emails that use similar words.

Description: Content is summarized by showing salient passages or by automatically summarizing threads. Identity is summarized by showing the bio, and interactive investigation of identity relationships is also supported by visual navigation of the communication graph using TreePlus (Lee et al., 2006). In the bio, connectence and centrality are reported as categorical values: "Medium" if within one standard deviation of the mean, "Low" if more than one standard deviation below the mean, and "High" if more than one standard deviation above the mean. Because the automatic name detection

accuracy was rather low (due to a lack of domain-specific training data), extracted names were displayed in the email content viewer, but they were not used as identity attributes.

Our initial NetLens-Email implementation, using MySQL with fully normalized relations, produced unacceptable response times when scaled up to the full CMU version of the Enron collection (which is two orders of magnitude larger than the Berkeley subset). In future work we expect to overcome this limitation by selectively indexing and selectively denormalizing some relations, but we adopted a simpler expedient for the design walkthroughs described in this article. The key idea was to use a two step process in which the user first sees an overview of entire email collection with a simplified version of NetLens that supports only selection of a subset of the collection that is then exported to a fully functional NetLens interface for further investigation.

4.4 The User Experience: An Example of Rich Interaction

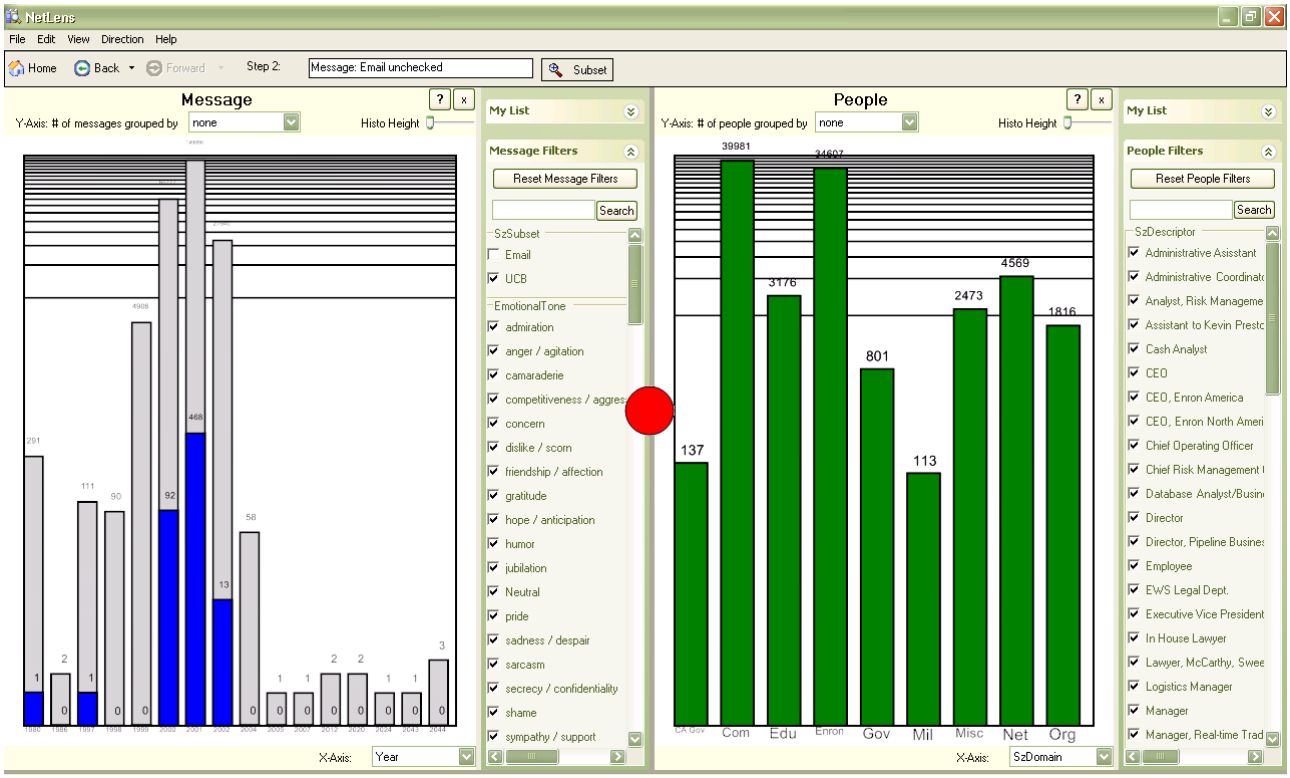
Imagine that we want to explore email between members of the senior management team at Enron. We might begin with several broad questions such as: "which executives relied most heavily on email?", "what kinds of topics did they discuss?" and "are unusual patterns evident in the email activity?"

To illustrate how NetLens email could be used to explore such issues, we start by restricting the content to the 1,700 annotated messages in the Berkeley subset (Figure 3(a)). We first identify the Enron employees among the senders and recipients of email in that subset by selecting "Domain" (i.e., the domain-name part of the email address) on the X-axis of the right (people) side overview (Figure 3(b)). Selecting the Enron (i.e., @enron.com) bar in the resulting histogram yields in a filtered list of names and titles for Enron employees. We then restructure the right (people) side overview to show the distribution by title, noting 17 distinct titles (Figure 3(c)). Deselecting titles that are not of interest using the checkboxes in the filter region on the right (people) side further filters the list to 27 email addresses. To see the emails sent by those people, we project that set of people to the left (content) side by with a right-click on the solid oval, followed by selection from the pull-down menu of a left arrow with "Sent by" semantics (Figure 3(d)).

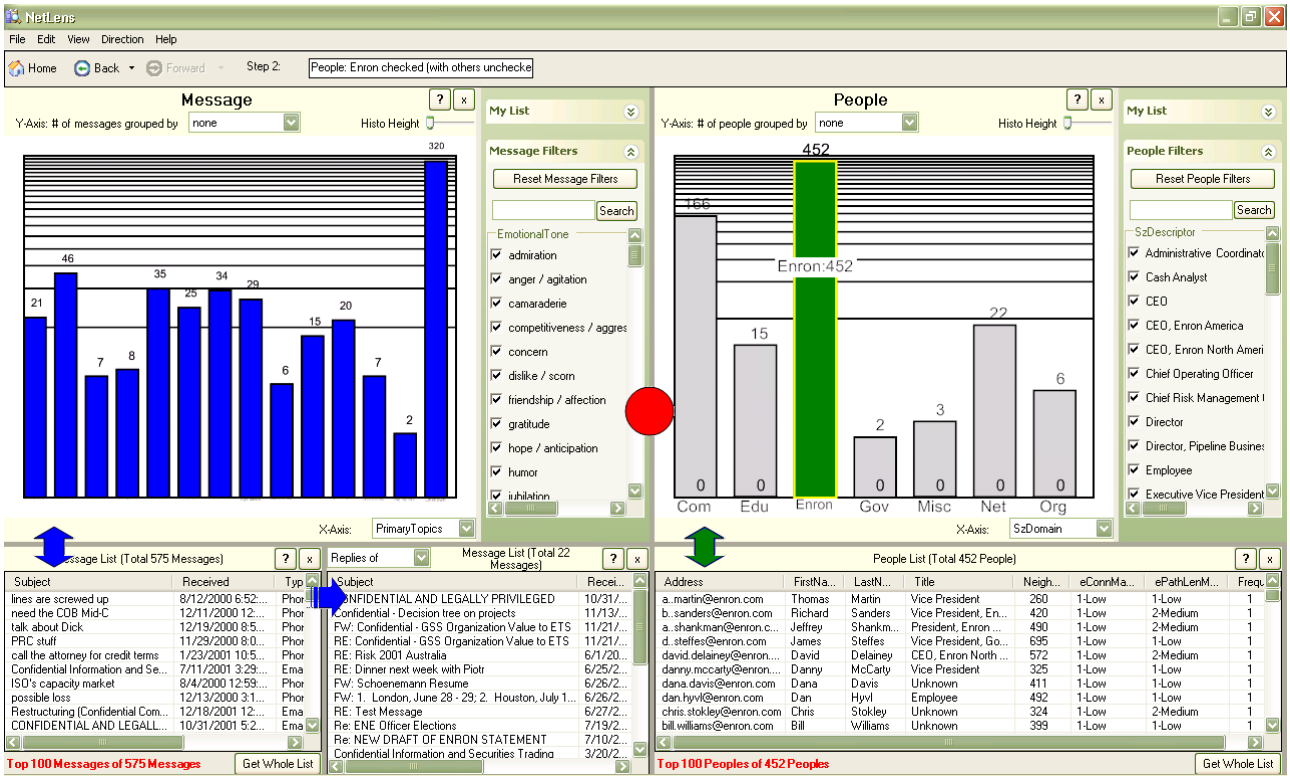
We now see just 25 of the emails sent by those 27 people are included in this 1,700-message subset. By changing the email overview to sort by "emotional tone" (a UC Berkeley annotation

category), we can see that one-third of those emails are annotated as confidential (Figure 3(e)). Restructuring the left (content) side overview to display the distribution by the time of transmission (with one hour granularity), we observe that some were sent well after the end of a normal workday. Selecting the histogram bars between 8:00 PM and 3:00 AM results in retaining the 12 messages shown in the leftmost item list on the left (content) side (Figure 3(f)). In this example, the rightmost item list on the left (content) side shows additional messages that are present in the same thread as at least one message in the leftmost idem list side; selecting any message in the leftmost item list would resort the rightmost item list to display other messages from that same thread first (with highlighting; not shown). Double-clicking on the first email in the leftmost item sets on the left (content) side brings up a popup window that shows the content of an individual email (the email viewer also lists names of people mentioned in that email on the right side of the popup window) (Figure 3(g)).

Reversing the direction of the projection arrow and selecting "Sent to" semantics for the projection operation shows who received those confidential late-night emails. There are 34 recipients, of whom just 8 have an Enron email address (Figure 3(h)). We could continue our exploration by viewing biographies for the non-Enron people who received confidential late-night messages, but in the interest of brevity we terminate our first example at this point.

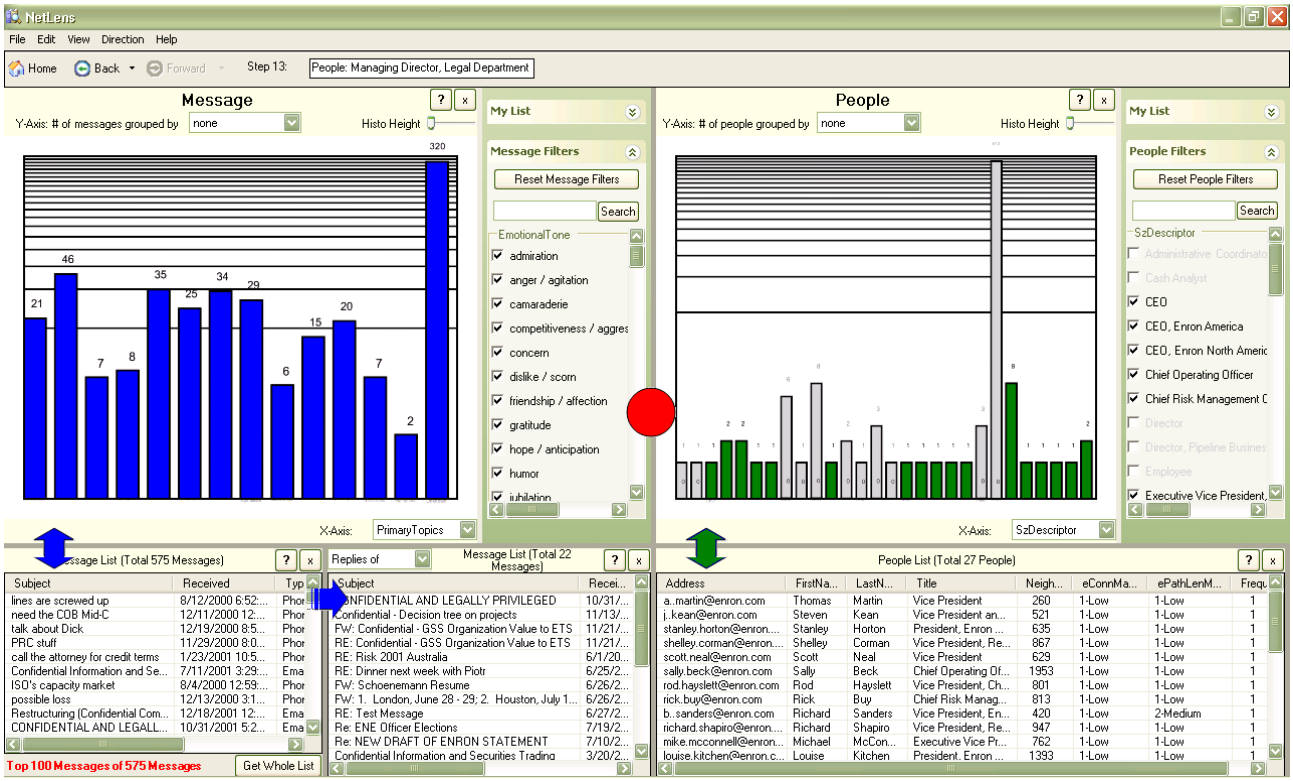


(3a) Simplified NetLens interface for email overview of 268,980 non-duplicate emails[8]

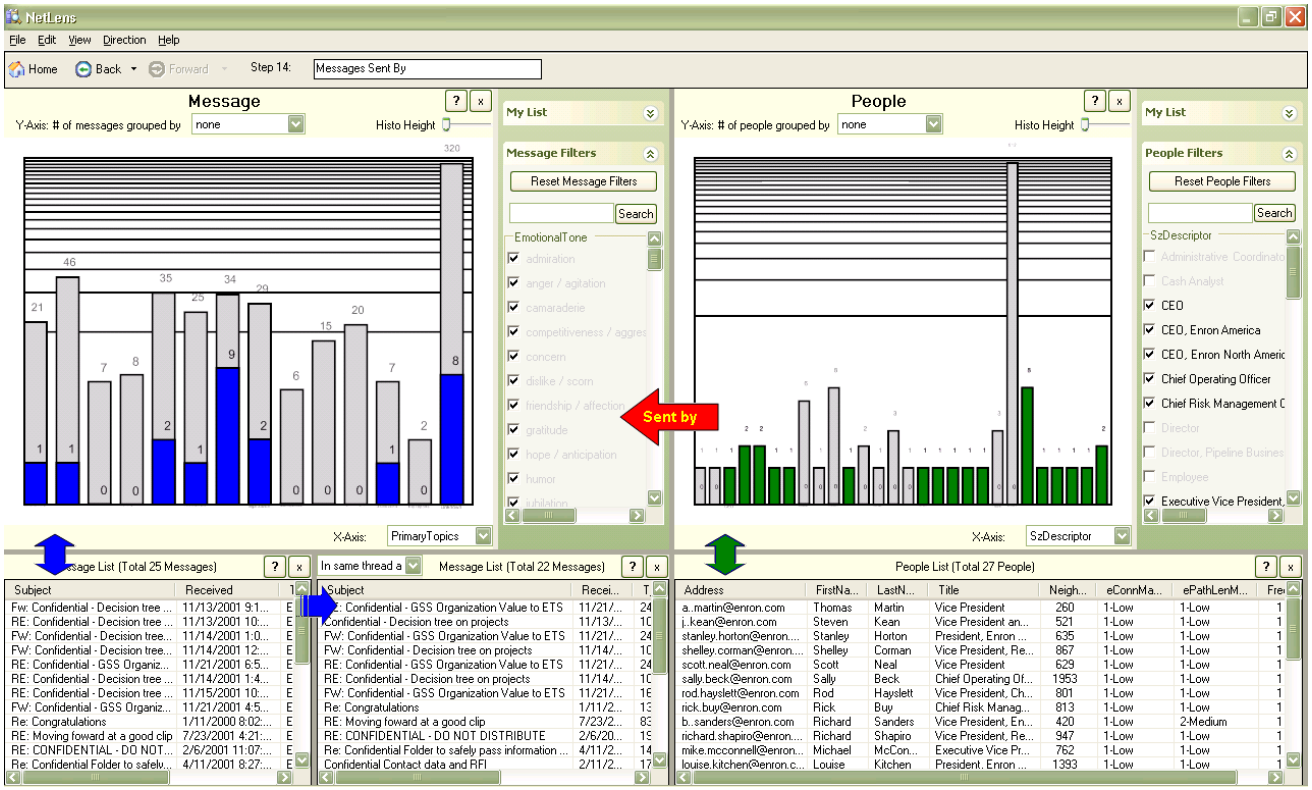


(3b) Select people who have Enron email addresses

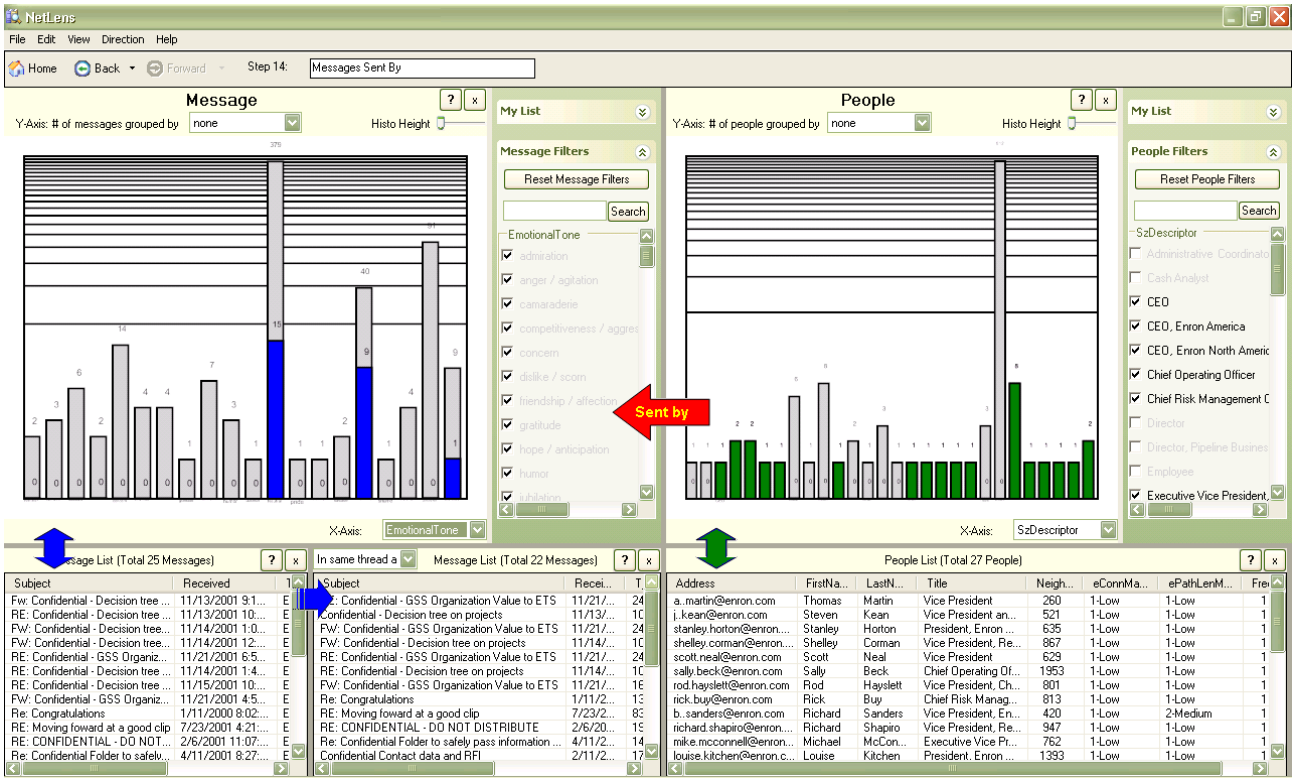
[8] All figures are available in higher resolution at www.cs.umd.edu/hcil/netlens/JASIST/figures.pdf



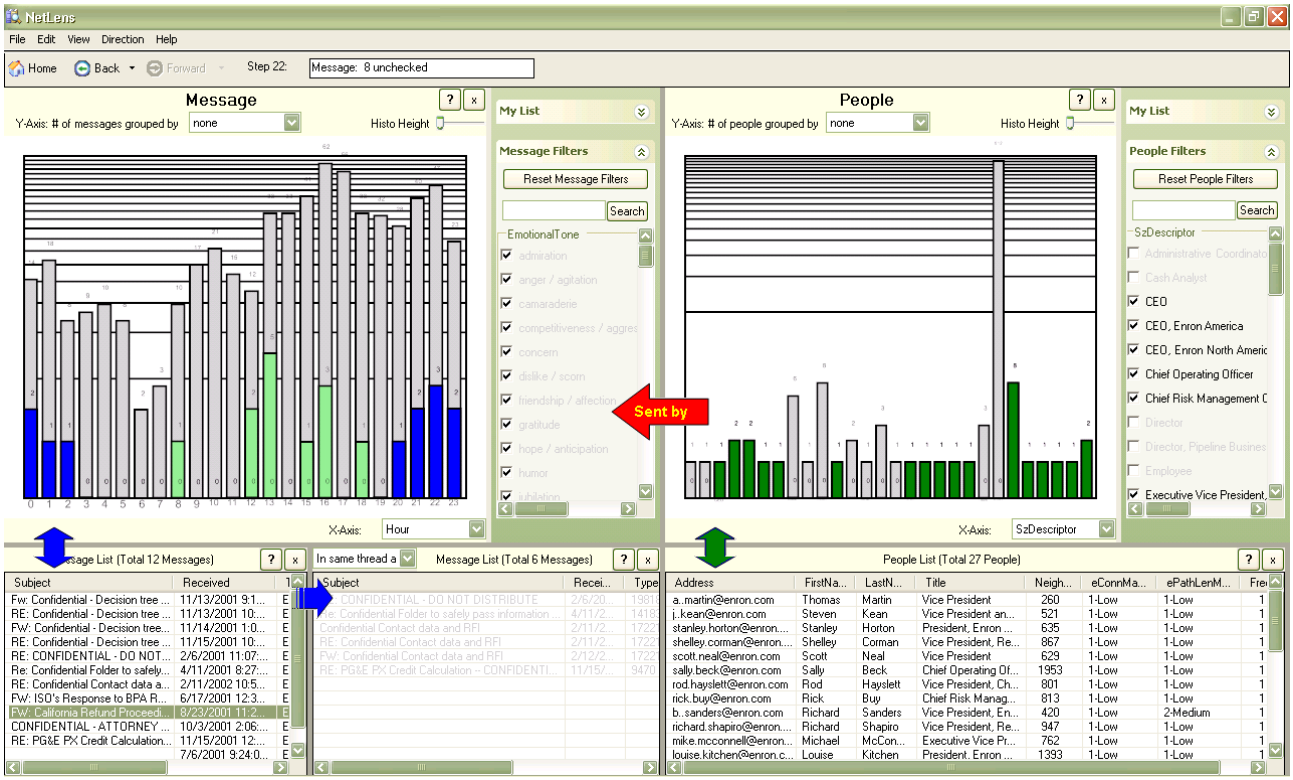
(3c) Change the distribution of people by their titles



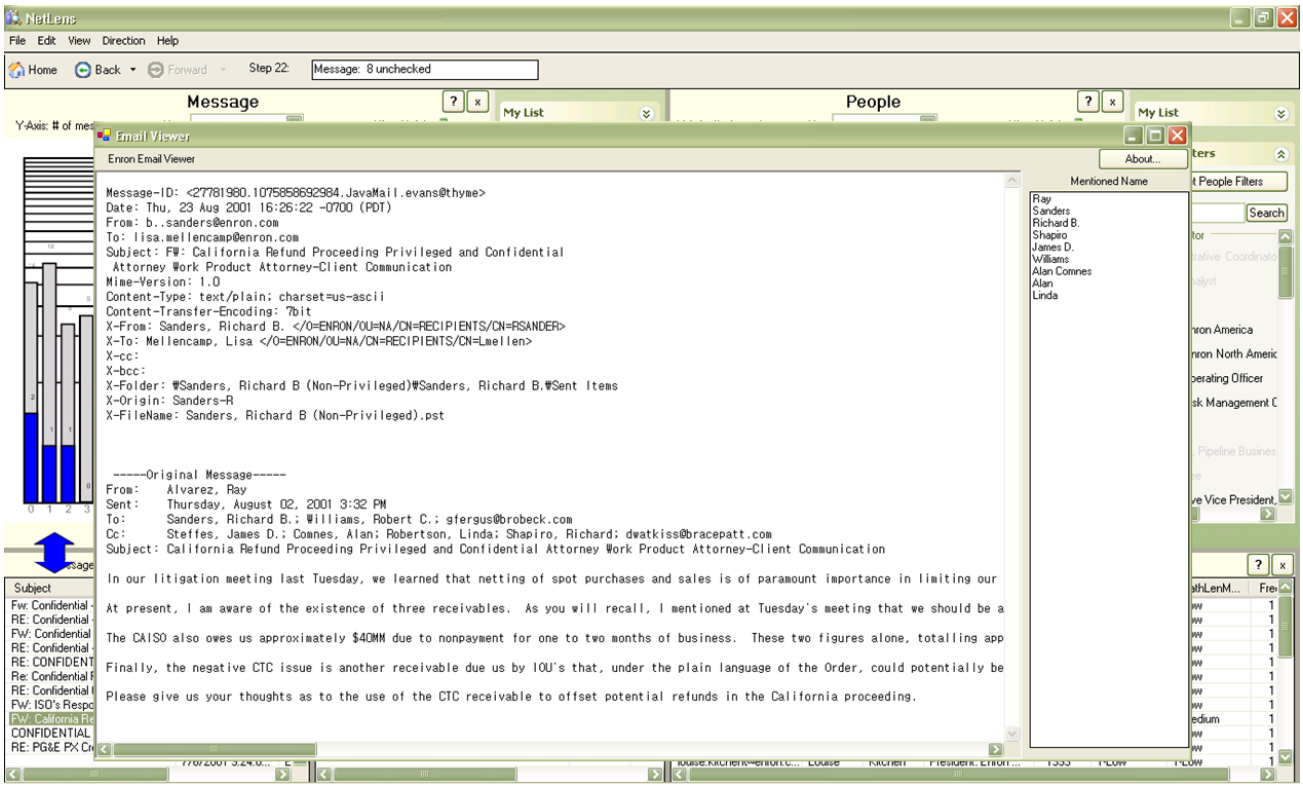
(3d) Find the emails sent by the selected people



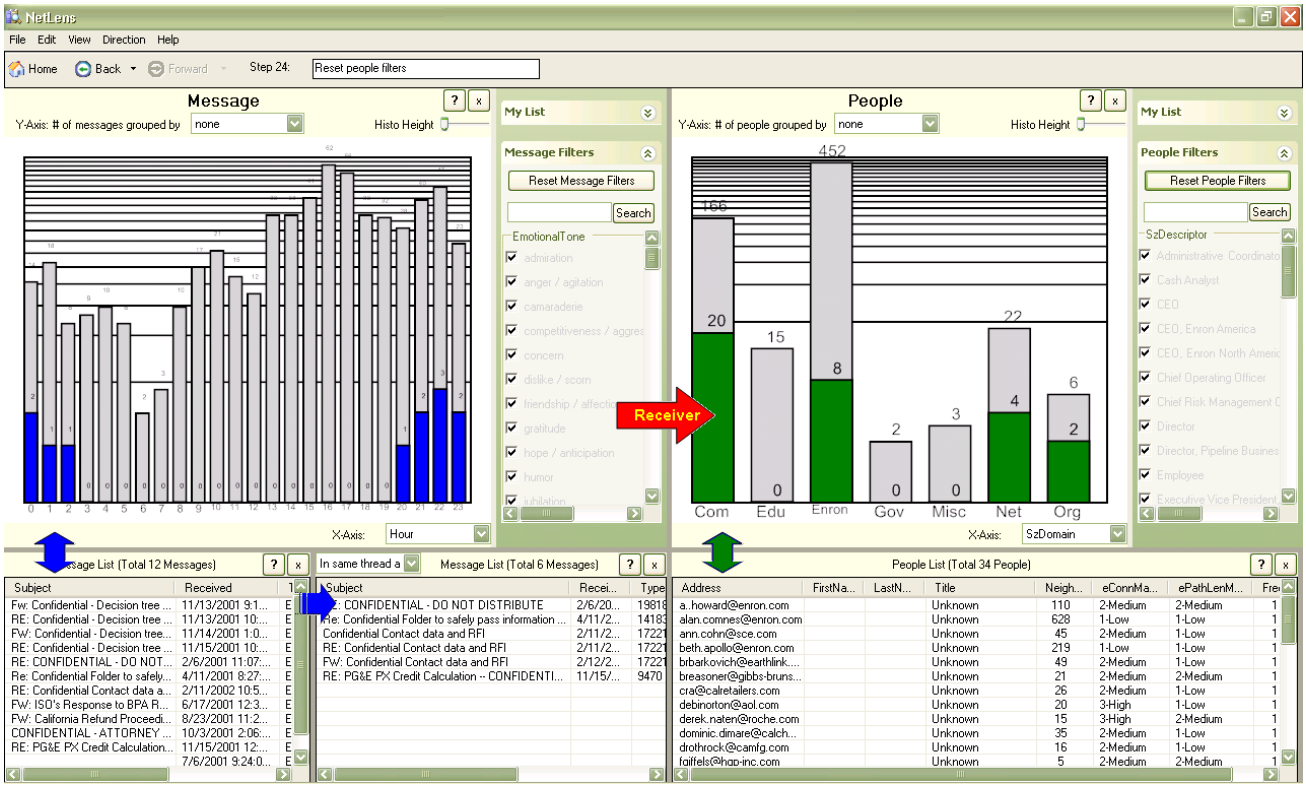
(3e) Change the distribution of emails by emotional tone



(3f) Select emails that were sent at unusual times



(3g) Show email body with mentioned names



(3h) Find the recipients of the selected emails

Figure 3: Illustrating use of NetLens to explore emails sent between Enron executives.

4.5 Another User Experience: An Example of Scalable Interaction

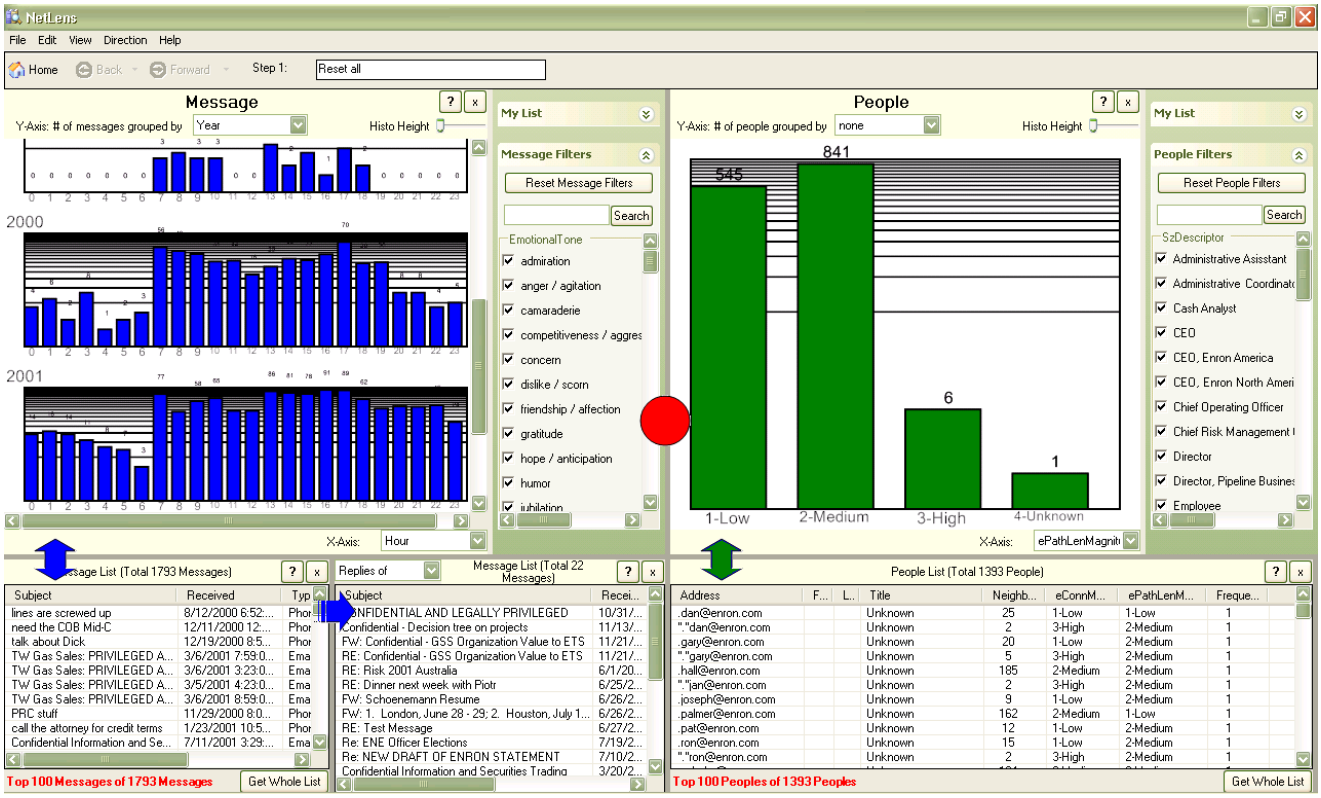
In this second scenario, we illustrate the use of NetLens-Email with the less sophisticated annotations that we have produced using fully automated techniques.

We start by looking at the overviews of the entire dataset and review the attributes available. We select the emails marked with emotional tone tag "secrecy" and export them to get better performance and faster interaction. We change the overview to see the distribution by year, then by hour of the day and notice a lot of messages in the middle of the night (Figure 4Figure 4(a)). We select them to see who sent those emails by selecting a projection operation for "Senders of Messages" (Figure 44(b)). We review the top of the list which is automatically sorted by number of emails sent. We select the top names and see the messages they sent by using the left arrow marked "Messages Sent By" (Figure 4(c)). The messages on the left are filtered and the overview shows the topic distributions showing a spike for "California". We select the bar for that topic, narrowing the emails to that topic, we review when the messages were sent and see that it spread several years but increased recently. We then see who the recipients of those emails were by using the projection operation for "Receivers of Messages".

We order the results by "path length magnitude" which brings the people that are "more connected to others" to the top of the list (the "about" button is available to see a definition of the social network statistics field (Figure 4(d)). This uncovers something interesting about two people mentioned in the emails: Tycholiz and Belden. We reset everything by pressing the "Home" button, and search for Tycholiz using the search box on the people side of the screen. Clicking on the name we can call up a bio (Figure 4(e)). Reviewing the various signatures helps us figure out Tycholiz's job or see what aliases he uses. We set the arrow to see the messages received by Tycholiz. We see a mixture of emails and phone calls which are rare, so we filter to see only the calls, and listen to them (Figure 4(f)). While listening we can see a list of people, places and keywords mentioned in each calls. We then switch to the emails. Some are part of a thread (indicated in the table with the number of emails in the thread) so we select the "thread summary" menu to generate a thread summary (Figure 4(g)). For a particularly interesting thread, we use the "find similar emails" menu to find other emails that use similar vocabulary but where not retained by the series of filters we used. Those emails are saved in "my list". This leads to more people of

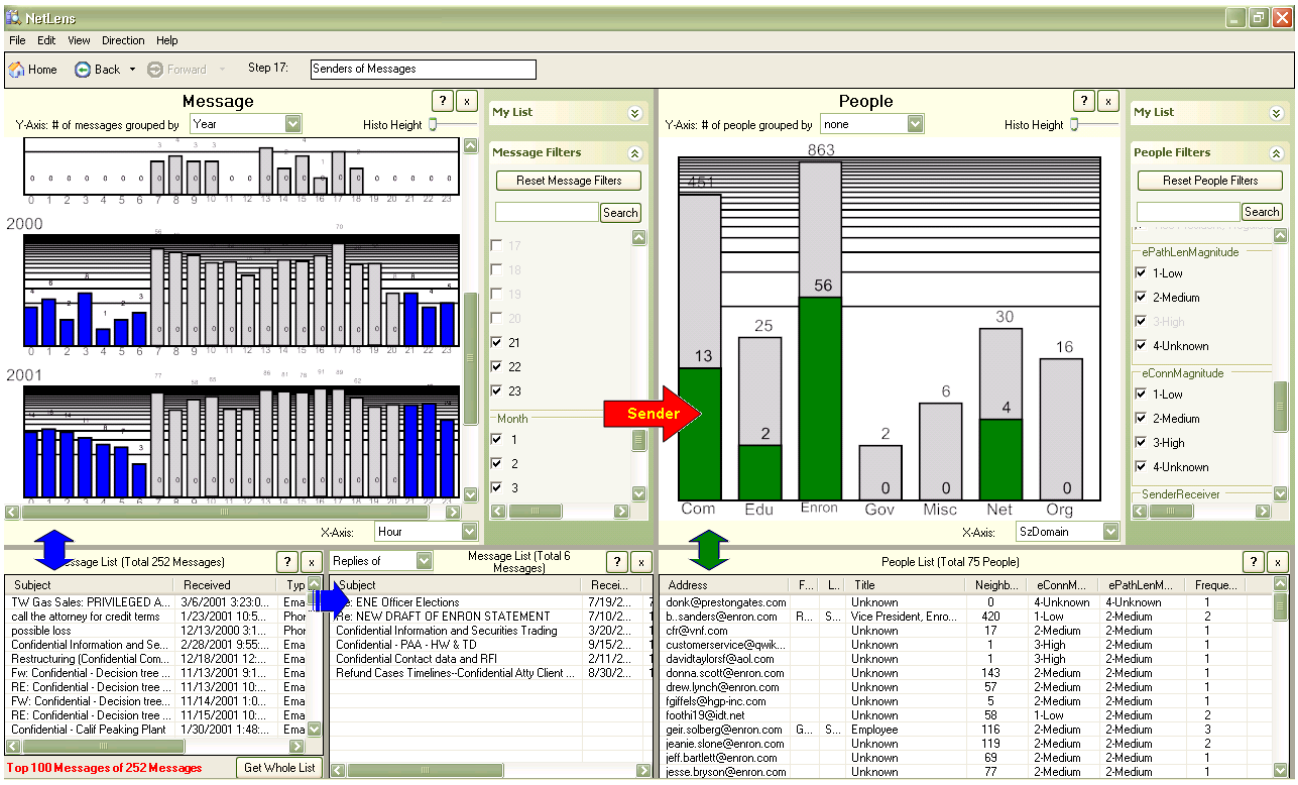
interest. We see a Treeplus view of the connection between Belden and the suspicious people we found (Figure 4(h)).

After all that work, we can review the history list of all the action we took. A name was generated automatically for each step but we can rename the steps to be more meaningful and save the history of work which was performed for audit trail or to send to a colleague. We can also simply retrace the steps at anytime, or jump to any step to start a new exploration.

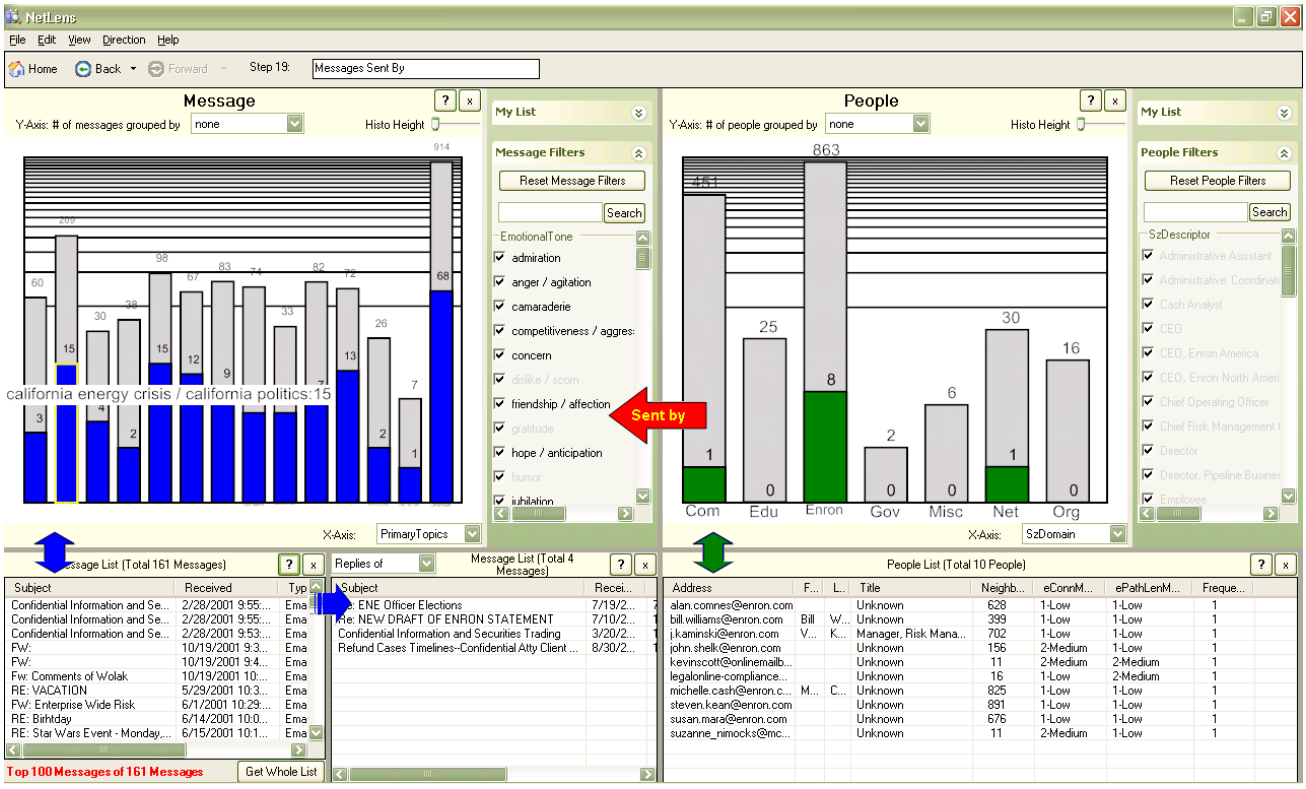


(4a) Overview of the messages selected from the entire emails[9]

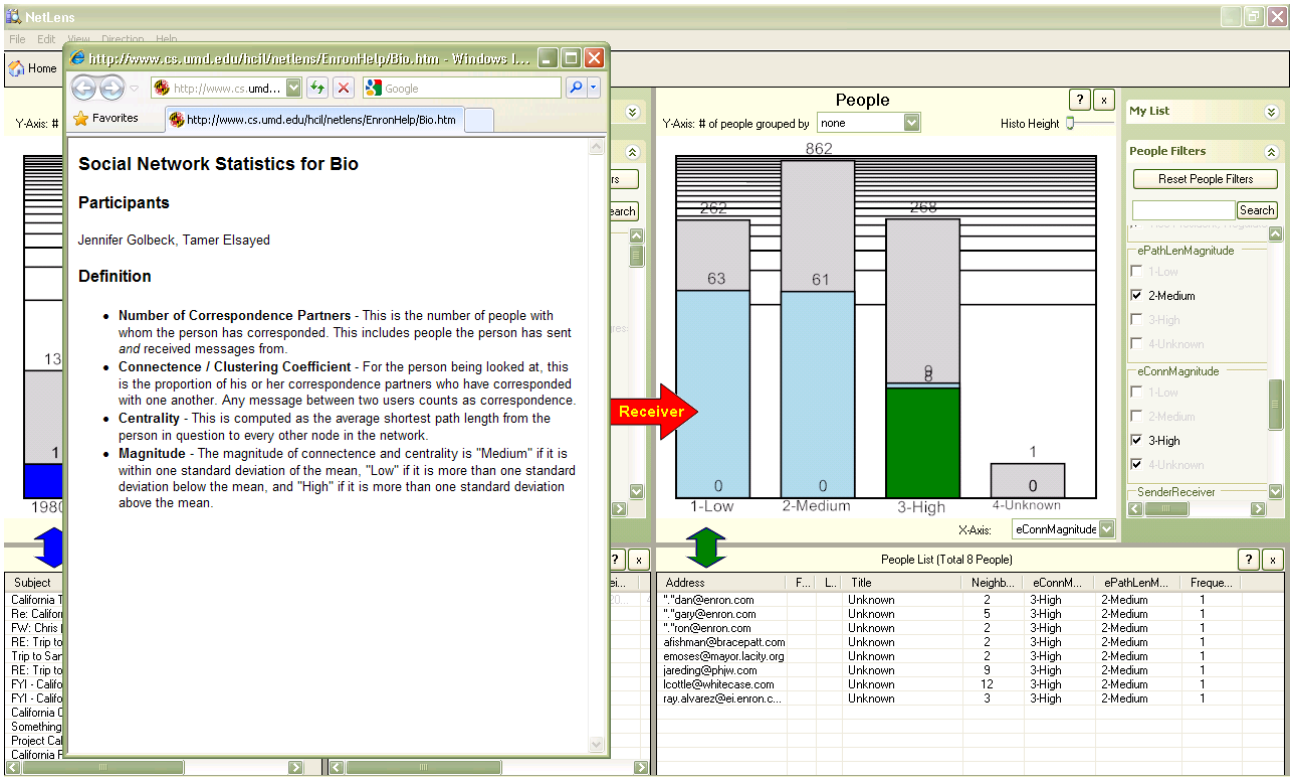
[9] Figures are available in higher resolution at http://www.cs.umd.edu/hcil/netlens/JASIST/figures.pdf



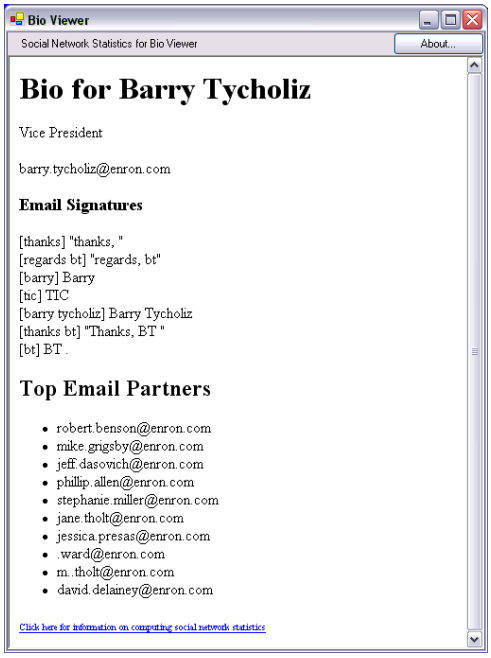
(4b) Select emails sent unusual time and find who received those emails



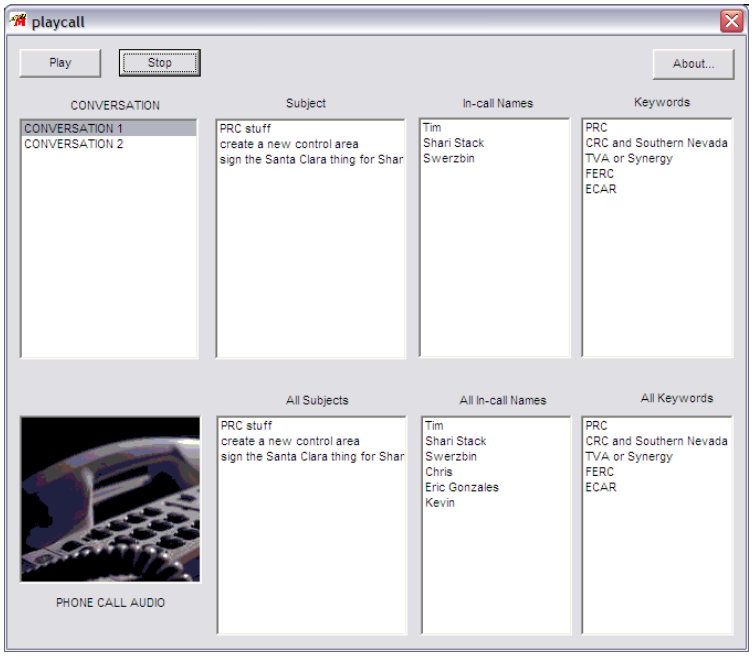
(4c) Select the frequent receivers of those unusual emails and see the topics of the emails



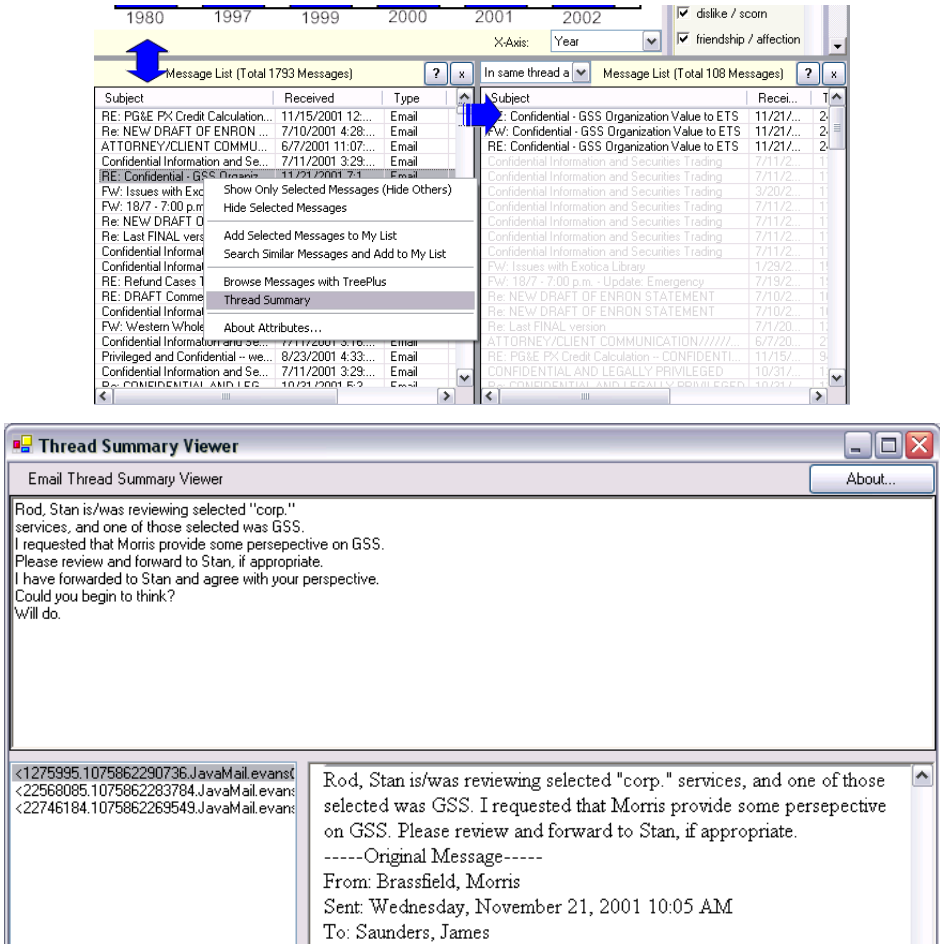
(4d) Use social network statistics to filter out more recipients



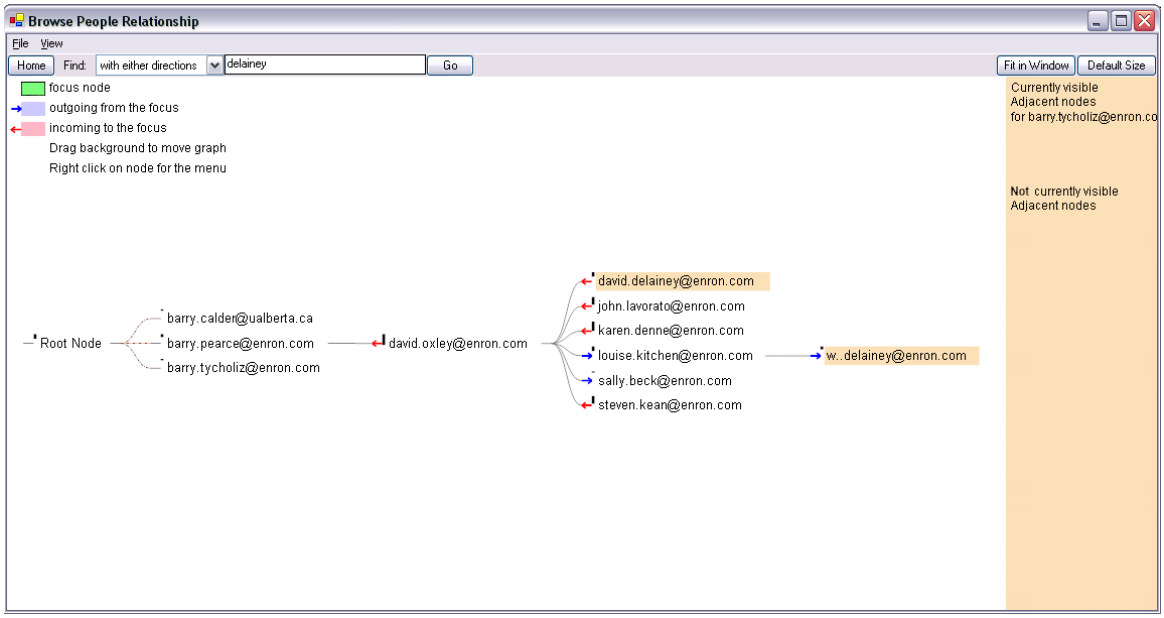
(4e) Bio of the selected person in NetLens



(4f) Interface integrated in NetLens to play phone calls



(4g) Thread summary generated by using the "thread summary" menu from the list of emails



(4h) Use TreePlus view to see the connection between the two people

Figure 4. Another scenario to illustrate use of NetLens to explore emails

5. Establishing a Baseline

Ranked retrieval systems such as Google Groups[10] are now a common access method for threaded discussion groups such as Usenet, and ranked retrieval has also been used for personal information management with large email collections (Dumais et al., 2003; Fisher et al., 2007). We therefore chose a ranked retrieval system tailored to email as a baseline to which NetLens-Email could be contrasted. We call this ranked retrieval system the Email Archive Explorer (EAE). EAE and NetLens-Email both provide access to the same collection and they share some common analytic components (both systems use the same header analysis, threading, and free-text search; only NetLens-Email uses the other preprocessing results).

The graphical interface shown in Figure 5 was principally designed to support exploration of the Enron collection by system developers rather than end users. The user can search in headers (by email address and/or name), the subject lines, new body text, quoted text, or any combination of those fields. For each search field, one or more query terms can be specified. The system can (at the user's option) respond to a query by presenting either a ranked list of individual emails or a ranked list of email threads. The list can be sorted in decreasing order of matching score (i.e., the usual "relevance ranking") or chronologically. Users can then select individual emails (after first selecting the thread containing that email, if using thread retrieval) for display in either its native format or in a color-coded format that highlights system-recognized components such as the main header, quoted headers, salutation line, and signature blocks. Nested levels of quotation can also be color-coded to help with visualization of the email content and structure. Query construction is a stateless process (i.e., a new search query simply replaces the previous query, and new search results are generated that bear no necessary relation to the previous results).

[10] http://groups.google.com

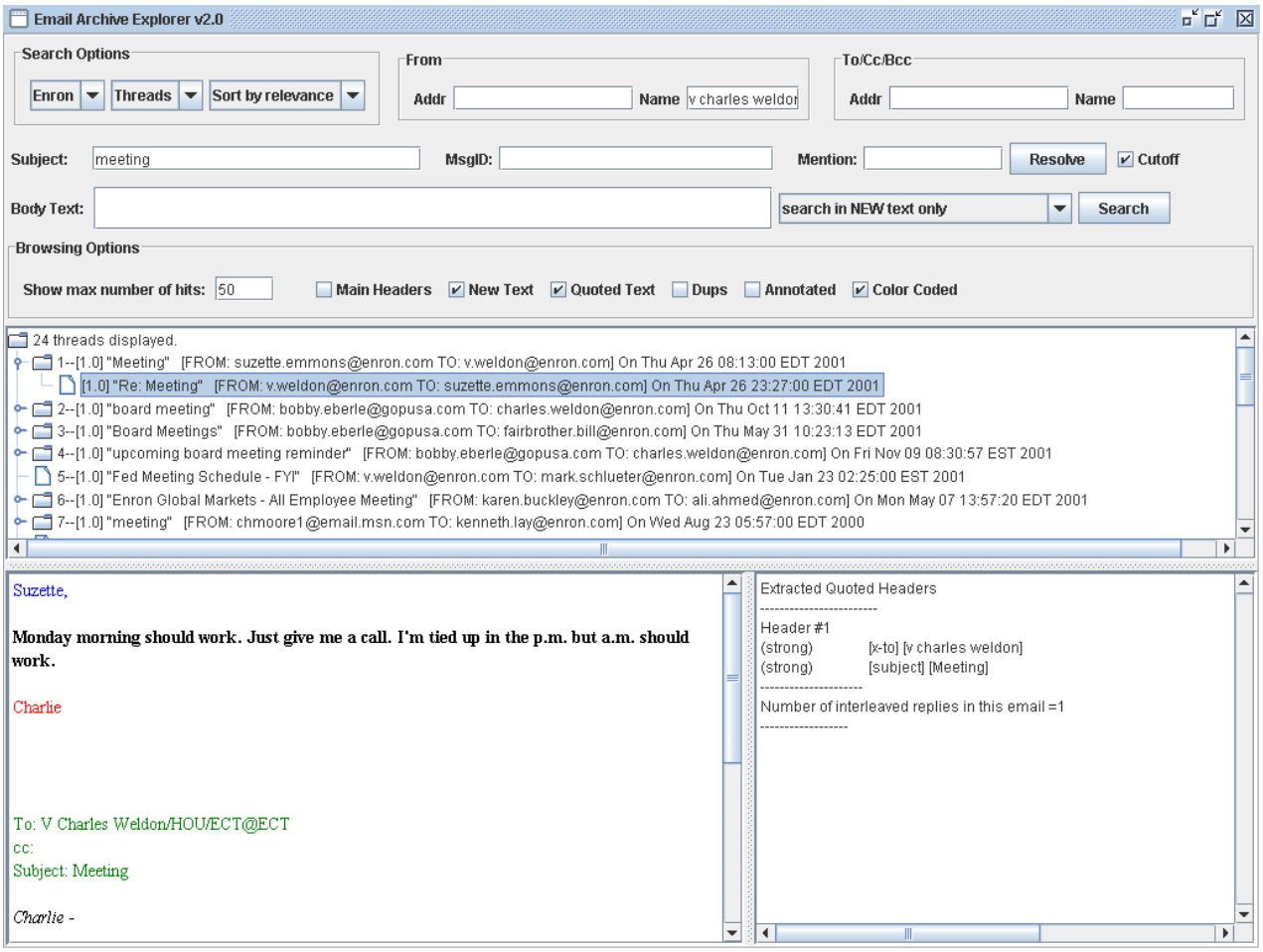


Figure 5. A screen shot of the Email Archive Explorer (EAE)

5. Expert Walkthroughs

Evaluating complex systems can be challenging, especially for exploratory tasks (White et al., 2006) and information visualization (Plaisant et al., 2008). Several evaluation approaches are possible:

1) *Flexibility* can be evaluated by assessing the time and effort that are required to adapt NetLens to new applications. The NetLens-Email system is one of three applications that we have developed to date (the third being a version for exploring photograph collections), so the details provided above contribute to our understanding of this aspect.

2) *Functionality* can be evaluated by assessing the degree to which the capabilities needed to perform a task are provided. The expert walkthroughs reported in this section begin to address this issue.

3) *Usability* evaluation relies on structured or unstructured user studies to measure how well people are able to understand and employ the system. Results from a small study reporting on use of NetLens for bibliographic information exploration are reported in (Kang et al., 2006).

4) Once a stable and scalable system is ready for operational employment, the overall *utility* of NetLens could be assessed with more complex user studies. Tightly controlled quantitative studies in laboratory settings offer some insight, but for exploratory tasks it is can be particularly useful to supplement quantitative studies with ethnographic observational studies of users doing their own work (Shneiderman and Plaisant, 2006).

Those four approaches form a natural sequence, with our focus in this article being flexibility and functionality, and our focus in this section being functionality. Drawing on widely used practices from software engineering, Polson et al. (1992) proposed goal-oriented "cognitive walkthroughs" of early "on-paper" designs as one way of exploring the degree to which envisioned systems would support a user's cognitive processes. Walkthroughs can continue as part of an iterative process for requirements elicitation as system components become available, initially with a focus on characterizing the required functionality (Connell and Shaffer, 1995). As designs mature, the balance between functionality and usability in these walkthroughs naturally shifts more towards usability, but at this stage in our work our walkthroughs focus principally on functionality. In this section we describe walkthroughs that we conducted with two domain experts.

We sought domain experts who could bring experience with similar sense-making tasks and an understanding of the challenges involved when working with large email collections. NetLens-Email was developed by one group within a large project team, and we found the needed expertise among other members of that team. One expert was a professional analyst with a Master of Library Science degree who had studied the use of Enron email in regulatory actions and litigation.

Our second expert was a professional archivist with a Ph.D. in Library Science who had served in an advisory role for a project in which long-term preservation of and access to a large collection of corporate email was a central issue.

Two sessions were conducted independently, one with each expert. At the outset of the session, the expert was asked to work with the two systems on a realistic open-ended exploratory task of their own choosing and to provide us with feedback on the design of our systems and with suggestions for our future work on tools to support exploration of large email collections. Both experts had previously used our EAE system, and we began our session with a 30-minute refresher session with that system in order to provide a point of comparison that was fresh in their mind. They then switched to NetLens-Email for the remaining 90 minutes of their session.

Before each expert started using each system we gave them a short demonstration of the available features of that system. We pointed out some known usability problems, and asked them to focus their attention principally on functionality rather than usability. They then used the tools themselves for the remainder of the session, although we answered questions about interface widgets and similar details when necessary so they could focus on higher-level tasks. Not surprisingly, both experts nevertheless made a substantial number of suggestions about usability (e.g., to somehow mark emails that had already been read) which we do not focus on here. In this section, we summarize the suggestions regarding functionality for each expert in turn, and then conclude by summarizing a few issues on which both experts commented.

A Professional Analyst

The first expert chose to work on the "401(k) question" (i.e., how Enron allegedly mishandled employees retirement funds). In the first part of the walkthrough, using the EAE system, this expert used complex Boolean queries (i.e., with AND and OR operators) and sorted by date to find an email from Enron's Chairman (Kenneth Lay) about the resignation of Enron's Chief Executive Officer (Jeffrey Skilling) in which Lay observed that Enron stock was "tanking." This led to a new question for our expert: did Lay say somewhere that his employees were being encouraged to buy Enron stock at that time? A search on Lay and Skilling led to email from

employees describing worries about stock prices. The EAE system's ability to quickly switch between levels of detail (e.g., showing only headers, only new text, or all the text, including the quoted text) proved to be particularly useful for this task.

That first expert then switched to NetLens. The multiple overviews were used to examine patterns in the results before starting to read the emails. For example, after a search on "401k" the expert observed that those emails were limited to 2000-2002 and saw an interesting pattern of daytime use of email in 2000 and nighttime use of email in 2001. The overviews proved to be less useful when looking at attributes that were less well populated (e.g., a distribution by job title proved not to be useful since that data was missing for most people). This expert wanted ways to create their own categories, and to correct erroneous metadata (e.g., adding additional job titles that the user could infer from emails that they read). This highlights the general issue of combining exploration with annotation, a capability that could be useful in many applications. The search and history features were used extensively, and the "my list" feature that we provided in which important emails could be saved was also remarked on positively.

Although the metadata about the emails and people was limited, this expert found that the available metadata sometimes helped them form a new question. For example, only 12 people had been marked as energy traders (which was Enron's core business), so finding one in a result set lead to a question focused on those traders. Filtering on emails sent by all traders followed by a search on "stock" in 2001 led to 15 emails, including one from Jim Schieger saying that he "has lost all confidence in management." This expert remarked, however, that an understanding of the provenance of the metadata would be important to them (e.g., how was a name attached to an email address, or why was a particular email tagged as "confidential?").

Our first expert was able to chain multiple search, filter, and selection steps together to construct complex queries by projecting back and forth between the content and people sides of NetLens-Email. After a while, however, it became clear that this expert—who was of course still novice at using the NetLens-Email system—had lost track of the state of the system. When we discussed that problem at the end of the session, this expert suggested that a "storyboard" approach might be preferable to our present "butterfly or flip flop" design using two fixed panels. The storyboard

they envisioned would create a new panel (for content or for people, as appropriate) to the right of the current panel whenever a projection operation is performed; the resulting strip of panels would automatically slide to the left to that two panels were still displayed, with the flow always left to right. An animation effect could visually indicate the transfer of data from one panel to the other, and a panel thumbnail browser could be used to generate a visual overview of history and state. This is actually an idea we had considered early in our design process, but we had chosen to first try a simpler fixed butterfly layout.

Overall, the first expert commented that even though NetLens-Email showed promise, it needed more work before it would be useful to analysts. They also observed that because NetLens-Email employs a novel interaction paradigm, it would be harder to learn to use it effectively.

A Professional Archivist

Our second expert encountered no difficulties with the EAE system; their comments during that part of the session focused on the processes that archivists use to explore collections (e.g., consulting multiple sources and taking notes about names and places) and not on system capabilities. When using NetLens-Email, this second expert explored many simpler questions rather than digging deeply into a single question as had our first expert. Chaining just two or three steps for each question resulted in disorientation being less of a problem in this case.

This expert made extensive use of pen and paper during the session, jotting down names and key terms (e.g., "blue dog"), and gave detailed comments on how they would use a variety of tools in parallel to conduct the task (e.g., perhaps searching the Web to learn what "blue dog" meant). They looked at automatically generated bios when they were available, commenting "this is just emails … historians will want to look at other things." These considerations led to a suggestion that we consider integrating NetLens with resources external resources such as glossaries. Adding information to external resources (e.g., those maintained by the researcher) was also mentioned as an important part of this expert's work process. In discussions after their session, this expert pointed out the potential value of names and time frames for linking to external resources. For

example, they suggested that a search on Ken Lay in the people panel of NetLens might also highlight his position on an organization chart.

The customizable overview region in NetLens-Email and the ability to filter an item set by selecting histogram bars were remarked on as helpful ("I like that, I can see where I am."). This expert commented that being able to follow a lead immediately (e.g., from an email, to people, to other emails those people sent) was useful, and it was an improvement over taking notes and then later performing a new search based on those notes. This expert drew a parallel between Netlens-Email overviews and the finding aids used in archives.

In general, this expert preferred threads over short messages, because "long emails are likely to be juicy." Showing more clearly which items are messages and which are threads—or even just making it easy to find long messages—was suggested as a possible improvement. Specifically, this expert suggested that additional attributes (e.g., presence in a thread, and message length) be made available for use in the overview panel.

Finally, this expert suggested that automatically assigning attributes that would seem to call for some degree of inference when assigned by a human annotator (e.g., "emotional tone") might not be seen as helpful by some users of archives since "an historian will only use hard facts, like dates or names."

Other Issues

Both of our experts expressed some frustration with the limitation of both EAE and NetLens-Email to maintaining a single state. In the case of NetLens-Email, this state evolves over time, but at any time only one state is available. Professional searchers are trained to construct partial result sets that can then be combined in various ways (e.g., set union, set intersection, and search within a set). The history maintained in the multi-panel functionality suggested by our first expert offers one way in which users might "roll back" to a previous state, but including more complex operators for working with results sets could add even greater complexity to what is already a rather complex interaction process.

Although both of our experts had some familiarity with the issues raised in the Enron investigations and the structure of the email collection that we used, they both encountered some complex technical or financial issues that simply could not be completely understood in a single two-hour session. This suggests that longitudinal studies with scholars pursuing their own research interests over an extended period might yield different types of insights. We are aware of one such community: researchers working with the 7-million-document Legacy Tobacco Document Library (which contains several hundred thousand emails). If NetLens can be extended to that scale (and in particular to work efficiently with the 1.5 TB of scanned documents in that collection), that could provide an excellent setting for our Multi-dimensional In-depth Longitudinal Case Study (MILCS) methodology (Shneiderman and Plaisant, 2006).

6. Conclusion

NetLens brings together three lines of research: user-directed exploration, representation based on content, and representation based on communication behavior. Each of these has been explored in depth individually, and there has also been considerable work on the three possible pairs (exploration of social networks, exploration of large document collections, and joint modeling of content and communication patterns). Where NetLens has the potential to make its strongest contribution is bringing all three together in a way that permits modular extension to new types of collections. NetLens-Email represents our first step in that direction. Through design experience and two functionality walkthroughs we have gained useful insights, some of which apply generally to the NetLens design (e.g., a storyboard format might help users remain oriented during complex tasks) and others of which are specific to NetLens-Email (e.g., signature blocks can be a useful element in a bio).

Two key lessons emerge from our study. First, the differences in work practice between analysts and archivists resulted in somewhat different conceptualization of the requirements for tools to support the sense-making task. Our analyst thought of what we had built as being in some sense new, providing capabilities that weren't available before. Our archivist, by contrast, saw what we had done as in essence replacing some, but by no means all, of what they would otherwise do using pencil and paper. NetLens emerged from research on information visualization, and it has now met the real world of work practices. Clearly, the time is right to apply participatory design

in the next iteration of this requirements discovery process. Another key lesson is that NetLens-Email offers so many opportunities to assist the user with their task that "feature creep" becomes a risk.

What we hope to accomplish is to make exploration and sense-making a qualitatively different experience from use of a search engine, with or without pencil and paper. Scenario-based walkthroughs have the potential to capture visceral reactions in addition to feature-by-feature critiques, although because of the limited time available in each session we have probably just scratched the surface of that important topic. To the extent that a consistent message emerged on this point, it is that we now need to begin to embed some implicit guidance on ways to productively use the vast range of capabilities that we can provide. For example, we might give some "hidden functions" that presently reside in pop-up menus more prominent locations that permit direct manipulation, and we might label some of our displays in ways that more clearly suggest the types of interaction processes that users might employ.

Our work with manual annotations suggests that there is considerable potential for exploring ways of automating these types of richer annotations. For example, in future work position titles might be recovered automatically from signature blocks or inferred from other types of references to individual roles from within the collection. As we saw with automatically recognized references to person names, we will have to think carefully about how best to use the results of such a process, though, since more advanced extraction and inference tools will naturally make some mistakes.

While extending NetLens into NetLens-Email we had to address not only issues specific to email, but also general problems such as scalability. Scaling to millions or billions of records will always remain a challenge for visualization unless real-time aggregation is provided. The solution that we implemented to support our walkthroughs is only partially satisfactory, as it requires users to reduce the dataset first before gaining full interactive access to the data. Ultimately, very large datasets may well require that we simply think differently (Shneiderman, 2008). Similarly working with email made it clear that NetLens needed more varied types of summaries than the original bar charts. Biographies or thread summaries are examples of summaries that are tailored

for email, but ultimately we need to look to developing a general framework for generating summaries that could be applied to a variety of domains.

Although email has been our focus to date, similar functions could be provided for archived collections of instant messaging, meeting records, automatically transcribed telephone calls, and similar informal sources that are likely to make up an increasing share of the legacy that we leave to future generations. What makes email an interesting focus at this point in time is that email collections of substantial size, complexity, and importance are now available to support this research. By allowing us to focus on how people will make sense of informally produced content that they themselves had no role in creating, we have gained a glimpse into what we expect will be an important topic for future research on information access.

Acknowledgements

The authors would like to thank our two participants for their assistance with this study, Jennifer Golbeck for biography generation, David Zajic and Bonnie Dorr for summary generation, Andres Kwasinski for annotating the transcribed telephone calls, and TJ Rogers for data integration. This work has been supported in part by the Joint Institute for Knowledge Discovery at the University of Maryland.

References

Balog, K., Azzopardi, L. and de Rijke, M.. (2006) Formal Models for Expert Finding in Enterprise Corpora, in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43-50, Seattle.

Baron, J. and Thompson, P. (2007) The Search Problem Posed by Large Heterogeneous Data Sets in Litigation: Possible future approaches to research, in *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, pp141-147, Stanford.

Bederson, B., Grosjean, J. and Meyer J. (2004) Toolkit Design for Interactive Structured Graphics, *IEEE Transactions on Software Engineering*, 30(8)535-546.

Connell J. and Shaffer R. (1995) *Object-Oriented Rapid Prototyping*, Prentice-Hall.

Craswell, N., P. De Vries, A. and Soboroff, I. (2005) Overview of the TREC-2005 Enterprise Track, in *The Fourteenth Text Retrieval Conference*, Gaithersburg.

Culotta, A., Bekkerman, R. and McCallum, A. (2004) Extracting Social Networks and Contact Information from Email and the Web, in *Proceedings of the Conference on Email and Anti-Spam*, Mountain View.

de Vel, O. (2000) Mining E-mail Authorship, in *Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining*, Boston.

de Vel, O., Anderson, A., Corney, M. and Mohay, G. (2001) Mining E-mail Content for Author Identification Forensics, *SIGMOD Record*, 30(4)55-64.

Dervin, B. and Foreman-Wernet, L. (eds.) (2003) *Sense-Making Methodology Reader: Selected Writings of Brenda Dervin*, Hampton Press.

Diehl, C., Getoor, L. and Namata, G. (2006) Name Reference Resolution in Organizational Email Archives, in *Proceedings of the Sixth SIAM International Conference on Data Mining*, pp. 55-64, Bethesda.

Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R. and Robbins, D. C. (2003) Stuff I've Seen: A System for Personal Information Retrieval and Re-use, in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 72-79, Toronto..

Elsayed, T. and Oard, D. (2006) Modeling Identity in Archival Collections of Email: A Preliminary Study,in *Proceedings of the 3rd Conference on Email and Anti-Spam*, Mountain View.

Elsayed, T., Oard, D. and Namata G. (2008) Resolving Personal Names in Email Using Context Expansion, in *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics (ACL)*, pp. 941-949, Columbus, OH.

Fisher, D., Brush, A., Hogan, B, Smith, M. and Jacobs, A. (2007) Using Social Metadata in Email Triage: Lessons from the Field, in *Symposium on Human Interface*, pp. 13-22, Beijing.

Goldstein, J., Kwasinski, A., Kingsbury, P., Sabin, R. and McDowell, A. (2006) Annotating Subsets of the Enron Email Corpus, in *Proceedings of the Third Conference on Email and Anti-Spam*, Mountain View.

Heer, J., *Exploring Enron: Visual Data Mining of E-Mail*, Web site, 2005, http://jheer.org/enron. (visited November 15, 2007).

Hemminger, M. (2005) *The U.S. Federal Energy Regulatory Commission's 2002 Investigation of Enron: Information Seeking, Questions, Methods and Conclusions*, Technical Report CS-TR-4775, University of Maryland, College Park, September 2005.

Kang H., Plaisant C., Lee B. and Bederson B. (2007) NetLens: Iterative Exploration of Content-Actor Network Data, *Information Visualization*, 6(1)18-31.

Keila, P. and Skillicorn, D. (2005), Detecting Unusual and Deceptive Communication in E-mail, in *Proceedings of the 2005 Conference of the IBM Centre for Advanced Studies on Collaborative Research*, pp. 117-125, Toronto.

Kerr, B. (2003): Thread Arcs: An Email Thread Visualization, in *2003 IEEE Symposium on Information Visualization*, p. 27, Seattle.

Klimt, B. and Yang, Y. (2004) Introducing the Enron corpus, in *Third Conference on Email and Anti-Spam*, Mountain View.

Lee B., Czerwinski M., Robertson G. and Bederson B. (2005) Understanding Research Trends in Conferences Using PaperLens, in *Extended Abstracts of the SIGCHI Conference on Human Factors in Computing System*, pp.1969-1972, Portland.

Lee, B., Parr, C., Plaisant, C., Bederson, B., Veksler, V., Gray, W. and Kotfila, C. (2006) TreePlus: Interactive Exploration of Networks with Enhanced Tree Layouts, *IEEE Transactions on Visualization and Computer Graphics*, 12(6)1414-1426.

Leuski, A. (2004) Email is a Stage: Discovering People Roles From Email Archives, in *Proceedings of 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 502-503, Sheffield.

Leuski, A., Oard, D. and Bhagat, R. (2003) eArchivarius: Accessing Collections of Electronic Mail, in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 468, Toronto.

Lewis, D. and Knowles, K. (1997) Threading Electronic Mail—A Preliminary Study, *Information Processing and Management*, 33(2)209-217.

McArthur, R. and Bruza, P. (2003) Discovery of Implicit and Explicit Connections Between People using Email Utterance, in Proceedings of the Eighth European Conference of Computer-Supported Cooperative Work, pp. 21-40, Helsinki.

McCallum, A., Wang, X. and Corrada-Emmanuel, A. (2007) Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email, *Journal of Artificial Intelligence Research*, 30(2007)249-272.

Meho, L. and Tibbo, H., (2003) Modeling the Information Seeking Behavior of Social Scientists, *Journal of the American Society for Information Science and Technology*, 54(6)570-587.

Minkov, E., Cohen, W. and Ng, A. (2006) Contextual Search and Name Disambiguation in Email using Graphs, in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 27-34, Seattle.

Oard, D., Hedin, B., Tomlinson, S. and Baron, J. (2008) Overview of the TREC 2008 Legal Track, in *The Seventeenth Text Retrieval Conference*, Gaithersburg

Perer, A., Smith, M. (2006) Contrasting Portraits of Email Practices: Visual approaches to reflection and analysis, *Proc. of Advanced Visual Interfaces*, 389-395

Perer, A., Shneiderman, B. and Oard, D. W. (2006) Using Rhythms of Relationships to Understand Email Archives, *Journal of the American Society of Information Science & Technology 57*(14)1936-1948.

Perer, A., Shneiderman, B. (2005) Beyond Threads: Identifying Discussions in Email Archive. *IEEE Information Visualization*. 41 - 42

Plaisant, C., Fekete, J. D., Grinstein, G. (2008) Promoting Insight Based Evaluation of Visualizations: From Contest to Benchmark Repository, *IEEE Transactions on Visualization and Computer Graphics*, 14(1)120-134.

Polson, P., Lewis, C., Rieman, J. and Wharton, C. (1992) Cognitive Walkthroughs: A Method for Theory-Based Evaluation of User Interfaces, *International Journal of Man-Machine Studies*, 36(5)741-773.

Proulx, P., Tandon, S., Bodnar, A., Schroh, D., Harper, R. and Wright, W. (2006) Avian Flu Case Study with nSpace and GeoTime, in *IEEE Symposium on Visual Analytics Science and Technology*, pp. 27-34, Baltimore.

Rohall, S. L., Gruen, D., Moody, P., Wattenberg, M., Stern, M., Kerr, B., Stachel, B., Dave, K., Armes, R. and Wilcox, E. (2004) ReMail: A Reinvented Email Prototype, in *Proceedings of CHI 2004*, pp. 791-792, Vienna.

Sack, W. (2000) Discourse Diagrams: Interface Design for Very Large Scale Conversations, in *Proceedings of the 33rd Hawaii International Conference on System Sciences*, pp. 3034-3044, Maui.

Seo J. and Shneiderman B. (2005) A rank-by-feature framework for interactive exploration of multidimensional data, *Information Visualization*. 4(2)96-113.

Shneiderman, B. (2008) Extreme Visualization: Squeezing a Billion Records into a Million Pixels, *Proc. ACM SIGMOD 2008 Conference*, ACM, New York, 3-12

Shneiderman, B. and Plaisant, C. (2006) Strategies for Evaluating Information Visualization Tools: Multidimensional In-depth Long-term Case Studies, in *Proceedings of BELIV'06, BEyond time and errors: novel evaLuation methods for Information Visualization, a workshop of the AVI 2006 International Working Conference*, pp. 38-43, Venice.

Smith, M. and A. Fiore (2001) Visualization Components for Persistent Conversations, in *Proceedings of ACM CHI 2001*, pp. 136-143, Seattle.

Thomas, J., Cook, K. (Eds.) (2005) *Illuminating the Path, the Research and Development Agenda for Visual Analytics*, IEEE Press.

Tyler, J., Wilkinson, D. and Huberman, B. (2005) E-Mail as Spectroscopy: Automated Discovery of Community Structure within Organizations. *The Information Society*, 21(2)133-141.

Venolia, G. and Neustaedter, C. (2003) Understanding Sequence and Reply Relationships within Email Conversations: A Mixed-Model Visualization, in *Proceedings of ACM CHI 2003*, pp. 361-368, Fort Lauderdale.

Viegas, F., Boyd, D, Nguyen, D., Potter, J., Donath, J. (2004) Digital Artifacts for Remembering and Storytelling: PostHistory and Social Network Fragments, in *Proceedings of the 37th Hawaii International Conference on System Sciences*, pp. 40109-40109, Hawaii.

White R, Kules B, Drucker S, Schraefel MC. (2006) Supporting Exploratory Search, *Communications of the ACM*, 49(4)36-39.

Wise, J. A., Thomas, J., J., Pennock, K., Lantrip, D., Pottier, M., Schur, A. and Crow, V. (1995) Visualizing the non-visual: Spatial analysis and interaction with information from text documents, in *Proceedings of IEEE Information Visualization '95*, pp. 51–58, Atlanta.

Yeh, J. (2006) Email Thread Reassembly Using Similarity Matching, in *Proceedings of the 3rd Conference on Email and Anti-Spam*, Mountain View.

Zajic, D., Dorr, B. and Lin, J. (2008) Single-Document and Multi-Document Summarization Techniques for Email Threads Using Sentence Compression, in *Information Processing and Management*, 44(4)1600-1610.

Zhang, J. and Ackerman, M.S. (2005) Searching For Expertise in Social Networks: A Simulation of Potential Strategies, in *Proceedings of the 2005 International ACM SIGGROUP*, Sanibel Island.

Zheng, R., Li, J., Chen, H. and Huang, Z. (2006) A framework for authorship identification of online messages: Writing-style features and classification techniques, *Journal of the American Society for Information Science*, 57(3)378-393.