CommViz: Visualization of Semantic Patterns in Large Social Communication Networks

Karin Verspoor, Bahadorreza Ofoghi, Marlene Robles Granda

Abstract
We introduce CommViz, an information visualization tool that enables complex communication networks to be explored, exposing trends and patterns in the data at a glance. We adapt a visualization approach known as hive plots to reflect the semantic structure of the networks, a generalization we call semantic hive plots. The method efficiently organizes and provides insight into complex, high-dimensional communication data such as email or messages on social media. We present the architecture of the CommViz tool and its application to the Enron email corpus as a case study, demonstrating how the structure of the visualization enables investigation of patterns and relationships in a large set of messages. We also provide a user study performed with Amazon Mechanical Turk that shows the value of the tool for certain complex data interrogations, and further show how the incorporation of semantic structure on semantic coordinates can also be applied to parallel coordinates visualization. The integration of the social network characteristics with semantic attributes of the underlying data in a single visualization is, to our knowledge, a novel contribution of the work.

The tool can be accessed at http://commviz.eng.unimelb.edu.au
Code is available at https://bitbucket.org/readbiomed/commviz
The Enron email corpus is available from http://bailando.sims.berkeley.edu/enron_email.html

Keywords
information visualization, topic modeling, communication networks

Introduction
We present CommViz, an approach to supporting analysts who are struggling to gain insight into the patterns of communication in complex message datasets. As data sets increase in size and complexity, it becomes increasingly challenging to extract information from them. Visual representations of data can provide a means for users to analyse and reason about the content of data sets, making use of the power of the human perceptual system. Such representations are most effective when they are interpretable and reflect the analytic tasks that the analysts perform.

CommViz makes use of existing structure in social communication network data such as email, instant messaging, or on-line forums – utilizing metadata associated with each individual message indicating the sender, recipient(s), and the date and time of message transmission – and couples that structure with further contextual attributes, such as the inferred thematic content (topic) of the messages or the location of the sender. This allows an analyst using the tool to easily answer questions such as “Who sent a message to Whom in What context?” and to visually observe patterns in the data, such as hot topics, the most active communicators, and periods of intense message activity. It has applications in a range of contexts, including business intelligence, disaster management coordination, and law enforcement.

The tool encompasses several innovations:

- The adaptation of a visualization approach called hive plots, originally developed for biological network visualization, to enable clear expression of communication network structure. The novel contribution of this adaptation is to plot the data in terms of the data semantics rather than the quantitative characteristics of the network structure. We therefore call the adaptation semantic hive plots.

- The incorporation of an additional informative dimension beyond the \{sender, recipient, time\} metadata of a message, specifically the context of a message, which can also be used to reflect the content of the message. Thus the visualization captures social network structures including message directionality as well as message-specific context.

- Representation of all four attributes of a message in a single edge, while also supporting exploration of patterns involving different subsets of the attributes, across a large message collection. The approach generalizes from semantic hive plots to semantic coordinates, where axes are structured to capture two attributes simultaneously.

Dept. of Computing and Information Systems
The University of Melbourne

Corresponding author:
Karin Verspoor
Dept. of Computing and Information Systems
The University of Melbourne, Melbourne VIC 3010 Australia
Email: karin.verspoor@unimelb.edu.au
In this paper, we will introduce the characteristics of the semantic coordinates information visualization approach, its origins in hive plots, and its application in the CommViz tool to a large communication network. By adapting hive plots to this application, thousands of individual data points are represented and can be interrogated, but patterns in the data are nevertheless easily visually observable. The visualization is designed to directly reflect specific analytical questions of interest for investigating a communication network and therefore focuses on capturing the four core dimensions of the communication data \{sender, recipient, time, context\} in a single view.

The main data set we consider consists of emails sent between corporate users over a number of years, the Enron email corpus.\(^5\)\(^6\) To capture context in this data set, we apply a natural language processing technique called topic modeling\(^5\)\(^6\) to determine the thematic structure of the document collections. The Enron corpus itself has previously been a testbed for approaches to topic modeling.\(^7\) This research has demonstrated the value of thematic organization of large document collections and complex social networks. We will utilize the identified topics as a primary organizing element in the visualization.

We use this corpus as structurally representative of any large dataset consisting of users, with time-stamped textual messages being transmitted between them; there is nothing in the visualization that takes advantage of unique characteristics of email or the Enron corpus itself and therefore the visualization should be applicable to a range of communication networks where messages are exchanged among users.\(^4\) We support this claim by adapting it to two additional data sets. We will describe the pre-processing of the corpus to support creating the visualization, and present the overall architecture of the solution.

The research challenge that we address in this work is how to visually represent in a two dimensional, screen-based visualization the four core attributes of a single communication (message) for all messages in a large data set, while also providing insight into communication patterns at the network level. Put another way, we aim to design a visualization that exposes the network structure of user \{sender-recipient\} interactions while also integrating contextual and temporal attributes of those interactions. We will demonstrate that the tool can be used to elucidate meaningful patterns in the network corresponding to specific analytical questions of interest through a case study as well as a user study, and we will argue that the use of semantic hive plots has advantages over alternative approaches to visualising these communication networks.

Related work

CommViz draws on related work in network visualization. Its application to the Enron data builds on work in text analysis.

Network visualization

Networks have been the subject of many visualization projects, applied to data ranging from social networks (where nodes typically represent people, and links represent relations among them)\(^8\)\(^9\) to biological interaction networks (where nodes represent biological entities, and links represent their interaction in a biological process).\(^10\) A full review is beyond the scope of this paper; various surveys are available.\(^11\)\(^12\) A common representation of such node/link data sets is to represent nodes as circles and links as a line connecting two nodes that are related in some way. The placement of nodes in such visualizations will typically be in terms of force-directed algorithms.\(^13\)

Large data sets pose a challenge to these visualizations, as the visualization of a large number of nodes and links can result in a “hairball” or “spaghetti” effect in the visualization, with substantial overplotting, from which patterns in the connectivity structure are very difficult to discern.\(^2\)\(^14\) This problem can be observed for the Enron corpus in a representative social network visualization in Figure 9, which displays people (senders and recipients) of messages as nodes and includes links between two nodes when a message has been sent from one person to the other.

Other approaches have been proposed to address this challenge, including circular layouts, arc diagrams, and adjacency matrices, each with various pros and cons. Matrix-based layouts such as adjacency matrices have been argued to be preferred for dense networks, since they avoid all edge crossings and separate edges from nodes.\(^14\)\(^15\) These visualizations are applied primarily in the context of data with undirected, weighted relations between two objects (nodes), and there is an emphasis on utilizing link attributes to manage ordering within the matrices.\(^12\) However, the ordering of rows and columns can strongly influence interpretation.\(^14\) Recent work has extended 2-D adjacency matrices to allow for comparison across an additional dimension of time or some other attribute, e.g., through juxtaposed matrices\(^15\) and matrices that are “stacked” into a 3-D Matrix Cube.\(^16\)

Hive plots were introduced as a “rational” approach to visualizing large networks, using a parallel coordinate plot in which the axes are radially arranged, and in which network structural parameters such as a node’s clustering coefficient are used to assign nodes to axes, and then place each node along the axis in an interpretable coordinate position based on ordered structural values such as node degree or betweenness. The application of hive plots to networks has centered on organizing nodes in terms of structural elements of the network itself (i.e., quantitative network structure measurements).\(^2\)\(^17\) The resulting visualizations are argued to be quantitative, visually parsable and scaleable.

Visualization of message communications

A number of information visualization systems have been developed specifically for analysis of email, with an emphasis on the meta-data of email, including work that aims to identify communication structures within organizations,\(^18\)\(^19\)\(^20\) thread structures,\(^21\)\(^22\) or temporal patterns.\(^23\)\(^24\) Few visualization systems have been developed that consider the context or content of messages. Conversation Map was an early system that visualized the thematic structure of Usenet,\(^25\) providing three separate but connected visualizations of social networks, discussion themes, and lexical relationships. The Themail system is one example of a system that displays keywords arranged along a timeline;\(^26\) the visualization is
limited to showing the words corresponding to one sender-
recipient relationship at a time.

The recent TIARA system\textsuperscript{27} most closely resembles our
own work, in that it makes use of topic modelling to
summarize large text collections including email, and pro-
vides visualizations of metadata coordinated with the topic
visualizations. However, the topic-based visual summary
does not incorporate the social network relationships directly
or provide insight into the magnitude of communications
related to topics; our integration of both the social network
and the message themes in a single visualization is, to our
knowledge, a novel contribution of our work.

**Visualization of document topics**

Topic models are derived through the application of
probabilistic graphical models that infer sets of words
that cluster together in a document collection, in order
to discover themes or topics that are representative of
the documents in the collection.\textsuperscript{5} Because topic models
are probabilistic distributions over words, they are difficult
for people to inspect and verify. There have therefore
been efforts to facilitate organization and visualization of
topics. This work has primarily focused on corpus-level
analysis of topics, including visualization of the structure
and relationships of terms in topic models in TERMITE\textsuperscript{28}
with the aim of supporting assessment of the quality of topics
inferred. Other work, such as TopicMaps\textsuperscript{29} and TopicNets\textsuperscript{30}
has addressed visualizing the connectivity of topics and
documents throughout a collection. A number of other tools
such as Jigsaw\textsuperscript{31} and WordTree\textsuperscript{32} also aim to reveal patterns
in the text of document collections as a whole, although they
do not directly make use of topic modeling.

**Visualization of object attributes**

In CommViz, we aim to organize messages in terms of
their attributes in order to facilitate detection of meaningful
patterns relating to those attributes in the data, and we
therefore face the problem of visualizing high-dimensional
multivariate data. Methods such as Parallel Coordinates\textsuperscript{33}
and Adjacency Matrices can also be used to visualize such
multivariate data.

In Parallel Coordinates, individual objects/nodes are
positioned across several parallel vertical axes. Each vertical
axis corresponds to a specific dimension or attributes of
the data. Edges that connect the individual nodes among
the vertical axes represent the structural connectivity of the
different nodes and thus the respective dimension values.
First used in 1883 to visualize the ranks of the states
of the United States of America for several categories,\textsuperscript{34}
the mathematical foundations were systematically developed
into a coordinate system many years later by Inselberg.\textsuperscript{35}

Visualization of communication data, such as email corpora,
with a standard Parallel Coordinates approach can be
achieved by positioning individual data attributes on the
vertical axes. One possibility is to put senders, email topics,
timeframes, and recipients on the axes where a connecting
polyline (a line traced across all axes) would represent an
individual email between two individuals on a specific topic
in a given timeframe. However, finding trends related to
distant axes (such as those between sender and timeframe in
this case) will have to go through the middle axis (i.e., topic)
which will make it difficult to observe the sender/timeframe
patterns. The axes can be shuffled for the specific pattern
type under question, so in this particular case the two sender
and timeframe data dimensions can be made adjacent, but
patterns involving more than two dimensions are difficult
to discern. In addition, as pointed out by Jetnetzko et al.,\textsuperscript{36}
over-plotting of large data sets with Parallel Coordinates
may result in severe readability problems. We have also
found this when attempting to visualize the Enron email
corpus, as shown in Supplementary File 3. We therefore
propose an adaptation of Parallel Coordinates that builds on
the semantic coordinates inspired by our adaptation of hive
plots for semantic attributes in the Section Semantic Parallel
Coordinates.

Adjacency Matrices may also be used for visualization of
multivariate data. Data dimensions and respective
individuals can be mapped to the matrix dimensions and each
cell may then represent the connectivity of the individual
data points on the matrix dimensions. According to Liu et al.,\textsuperscript{37}
Adjacency Matrices, along with node-link diagrams,
are among the most popular visualization techniques for
network data visualization. Examples of visualizations with
Adjacency Matrices include MatrixExplorer,\textsuperscript{38} MatLink,\textsuperscript{39}
NodeTrix,\textsuperscript{40} and Cubix.\textsuperscript{41}

With a typical 2-dimensional Adjacency Matrix, it is only
possible to visualize communication data with respect to the
two dimensions mapped to the two matrix dimensions. In
addition, it may be possible to add a third data dimension
through cell color. With the methodology used in Cubix,\textsuperscript{41}
however, it is possible to add different slices to the
visualization to account for a fourth data dimension. For
instance, it may be possible to use sender and recipient
information as the two matrix dimensions, the cell color for
topics, and put such data of each timeframe in a separate
parallel matrix slice. As with Parallel Coordinates, such
visualization may lack clarity with respect to capturing
multi-dimensional data patterns when such patterns are
scattered on several disjoint matrix slices. In other words,
finding individual messages that are connected across the
different data dimensions (sender, topic, timeframe, and
recipient in this context) is not trivial with Adjacency
Matrices.\textsuperscript{42} In general, it is difficult to see the interaction
among two variables in these visualizations because the axes
would need to be sorted on two variables at once.\textsuperscript{43}

Our approach is similar to the Semantic Substrates\textsuperscript{12,44}
model, which uses categorical node attributes to organize
nodes into regions. In this model, non-overlapping regions
are defined to represent individual node attributes.

PivotGraphs\textsuperscript{45} use a grid-based layout combined with
aggregation of nodes based on shared values along a given
dimension to emphasize the patterns of connection between
groups of nodes. Because of the grid/matrix structure, they
are limited to analysing the relationships between two
dimensions at a time.

A recent proposal called DOSA\textsuperscript{45} (Detail to Overview
via Selections and Aggregations) allows both network
topology and multivariate data attributes to be visualized
simultaneously; it has been applied to the Enron data set
(among several others) to illustrate the value of the approach
for revealing communication patterns between and within
groups of users (e.g., Managers and non-Managers). In this approach, nodes can be projected into 2 dimensions ($x$ and $y$ coordinates) according to structural or categorical dimensions together, as related to the analytical question of interest. This is achieved by considering derived network structure values directly as data attributes. However, due to the use of an implicit grid to position nodes according to the value along a given dimension, the visual scalability is limited to a modest number (5-10) of attributes.

**CommViz visualization approach**

The visualization of social communications in CommViz is based on hive plots, an approach to laying out large networks based on a multi-dimensional coordinate system in which the axes are radially arranged rather than arranged in parallel. This allows connections between several axes to be visualised simultaneously.

**Hive plots**

Hive plots have been argued to be more interpretable than other representations, as well as reproducible and perceptually uniform. The primary applications of hive plots to network data have aimed to provide quantitative signatures of the data, and hence have utilized network properties such as connectivity or clustering as the basis of the coordinates, such that different networks can be compared structurally through the visualization.

The hive plot visualization benefits from several features that make it a preferred solution to other possible visualizations. First, hive plots can handle large datasets and large numbers of inter-relationships in an informative and reproducible manner. Second, the compact layout used in hive plots makes it straightforward to understand and visually quantify the connections among the elements in the visualization. Third, in contrast to other candidate visualization solutions, hive plots project the network of relationships on to a multidimensional coordinate system that enables organization and investigation of several distinct characteristics of the data simultaneously.

In hive plots, there are three main elements that can represent different types of information, namely axes, nodes, and edges or arcs. In addition, there is a linear organization of nodes along each axis that can also be used to convey information, that are referred to as segments of the axis.

**Adaptation to communication networks**

Our adaptation of hive plots to human communication networks centers on structuring the data in terms of its key semantic characteristics, such that the visualization supports inspection of the data in terms of several driving analytical questions. We call the adaptation semantic hive plots, because we make direct use of semantic characteristics of the communications to organize the visualization.

Specifically for messages, the visualization incorporates senders, recipients, contexts, and time into a single graphical representation. The primary focus question that can be investigated visually with the tool is “Who sent a message to Whom in what Context in a given Timeframe?” For textual messages, we can treat the topic or theme of the message as the key context; context=topic in this case. Other contexts are also possible, such as a location of the sender or an organizational unit.

We achieve this by organising two dimensions of information together along one primary axis, and the other two dimensions along the other primary axis. Then, a single link connecting one primary axis to the other represents one unique quadruple of \{sender, recipient, time, context\}. Trends involving 3-dimensional subsets of these, specifically \{sender, recipient, time\} independent of context (“Who sent a message to Whom in a given Timeframe?”), and \{sender, recipient, context\} independent of time (“Who sent a message to Whom in a given Context?”) are derived through the use of cloned axes, and links that go between a primary axis and its clone.

The core visual representation of CommViz consists of two primary axes, each representing individuals in the dataset in terms of one aspect of the focus question, context or time. For the focus Enron dataset, we treat context as the message topic, as determined by topic modeling.

One primary axis is divided linearly into segments representing topics (one topic per linear segment). Each segment contains a set of nodes representing all senders of email messages. For the Enron dataset, each segment contains a number of nodes equal to the total number of users in the Enron dataset (or subsample) and the axis has as many segments as topics. Each node on this axis represents a given sender who may have sent emails on a given topic. This axis is referred to as the user-topic axis.

The second primary axis is segmented linearly according to time, and hence has a segment for specific timeframes in the communication corpus. The year that a message was sent is the initial segmenting structure for the Enron data set along this axis. Each node in a segment represents a possible recipient of an email during a given timeframe. We call this axis the user-timeframe axis.

Links (arcs) connecting nodes in CommViz represent communications on a specific topic in a specific timeframe. Figure 1 depicts the basic structure of the visualization as applied to the Enron data, with an example arc between two users. The arc represents email communications (representing one or more messages) between the two Enron users about topic 2 in timeframe 2.

An important characteristic of this visualization structure is that the information along one axis carries over to the nodes on the other axis, via the connectivity of the edges. That is, the time information along the timeframe axis applies not only to the nodes on that axis, it also grounds the topics and the individual users linked on the other axis in that time period. This follows from the fact that a single link represents a quadruple of \{sender, recipient, timeframe, context\}. This property means that CommViz represents the four dimensions of the data in a visually efficient way.

To visualize the user interactions along one axis only (i.e., when not considering the information in another axis), each axis in the hive plot can be cloned. Figure 2 shows the same structure we have adopted for the Enron corpus when cloning has been utilized. Using this structure, it is possible to understand the interactions between two users either at a given time or about a given topic, exclusively.
Figure 1. Basic structure of the Hive Plot axes in CommViz, as applied to the Enron email data. It includes a user-topic axis and a user-timeframe axis. Along the user-topic axis, each segment corresponds to a topic and each node within the segment is a user who (may have) sent an email on that topic. Along the user-timeframe axis, each segment corresponds to a unit of time, with each node representing a user who (may have) received an email at that time. A link connecting a node on one axis to a node on another represents a set of emails from a given sender to a given recipient, on a given topic at a given time.

Figure 2. Cloned axes in the Hive Plot in CommViz. Cloning the user-topic axis results in links that capture emails between two users on a given topic, regardless of when the emails were sent. Similarly, by cloning the user-timeframe axis, we can represent emails between two users at a given time, regardless of the topic of the communication.

In the implementation (see Figure 3), we use color to reflect the number of emails that are represented by a link; the darker the link, the more emails are represented by the link, i.e., the “heavier” the interaction. A high density of links between the two user-topic axes within a given segment indicates that the topic is frequently discussed in the message corpus as a whole. A high density of links between the two user-timeframe axes within a given segment indicates many emails being exchanged during that timeframe.

We note that it has been previously observed that hive plots are by design flexible. Many heuristics or choices can be made in the hive plot model to map nodes to axes, and to position nodes along those axes. The difference in the context of our application is that we do not have a single static notion of what a node is, and our nodes do not correspond to a single real-world entity (e.g., a single message or a single user). Rather, the nodes plotted on the axes of the hive plot correspond to different subsets of the four message attributes that we aim to capture. The CommViz approach addresses the conceptual challenge of representing the four core attributes of an individual message in a single link in a 2-dimensional representation.

We will return to the visual analysis supported by CommViz in the use cases section.

Characteristics of CommViz

Here, we elaborate further on analyst-facing aspects of the visualization through the CommViz implementation. In addition to the core hive plots layout described above, we make use of information panels to provide further feedback to the analyst (see Figures 3-6). When observing the full Enron dataset through CommViz, the analyst will see summary data about the network in the top left panel (e.g., “Showing 9,252 links representing 56,210 emails between 90 people”). This panel will change to summary information about a given link in the network, when that link is in focus (due to hovering over it as the analyst mouses across the visualization) (see Figure 4). The summary captures the full details represented in the link – the sender, recipient, timeframe, topic, and number of the messages summarized in that link.

A further interactive panel provides both a legend for the color scheme of the segments along each axis (the topics for the user-topic axis, and the timeframes for the user-timeframe axis). In the case of the topics, as an analyst hovers over a topic label in the legend, more details of what the topic represents are provided (See Figure 6). In the current implementation of CommViz for the Enron data, these details are the set of words that make up the topic; however, some other representation of the topic could easily be substituted.
This legend also provides the entry point for zooming in to a subset of the data. Clicking on either a topic or a timeframe will drill down into the subset of data corresponding to that topic or timeframe. For instance, the analyst has zoomed into the year 2000 in Figure 5 and topic 19 in Figure 6. In response, the data used to populate the visualization is reorganized along the user-timeframe axis into sub-segments of the year; specifically it is split into the months of the year. Since there is no further logical breakdown of the topic itself, the data along the user-topic axis is not re-segmented, but all data not related to topic 19 in year 2000 is removed from the visualization. This allows the analyst to explore the emails within this subset of the data in more detail.

A final legend panel, seen in the top right of Figures 5 and 6, provides the analyst with an interpretation of the link color.

Representative application to email communication dataset

The Enron email corpus

The Enron email corpus was first made available to the public during the legal investigations by the Federal Energy Regulatory Commission (FERC) in May 2002. The dataset initially contained 619,446 email messages belonging to 158 Enron employees with an average of 757 messages per user. Most of the messages were sent between senior managers of the Enron Corporation. Each email object in the dataset contains the email address of the sender and recipient/s, message body, date, time, and subject. Attachments were excluded from the first publicly available Enron corpus; however, they can now be obtained from the Electronic Discovery Reference Model (EDRM). We exclude attachments for the purposes of this study.

A smaller version of the corpus has been made available online by William Cohen from Carnegie Mellon University. This version of the Enron corpus includes 517,431 emails from only 151 users after requests made from a number of affected Enron employees for their emails to be removed from the initial dataset. In addition, invalid email addresses in the initial dataset were replaced by user@enron.com when there was a recipient and the addresses with no recipients were converted to noaddress@enron.com. In this project, we made use of a relational database representation of Cohen’s version of the Enron corpus in MySQL. This database was created by Andrew Fiore and Jeff Heer at University California, Berkeley. A substantial amount of processing has been performed by the database creators to remove duplicates and normalize names; the final data set includes 89 individuals and 56,210 emails.

Topic modelling of Enron email corpus

Once the Enron corpus was ready for analysis, we applied a topic modelling technique to identify topics in the corpus as a whole, and to associate the content of each email message to one or more of those topics. The method of Latent Dirichlet Allocation (LDA), a standard topic modelling technique, was used. Topics in LDA are represented by different distributions over the list of terms in the vocabulary of the entire corpus. For instance, if the corpus only contained the three terms banana, book, and pen, then, banana (80%), book (7%), and pen (3%) would represent a fruit-related topic; whereas banana (5%), book (70%), and pen (25%) is representative of a book or stationary-related topic. If a document or message is considered to be mostly in the former fruit-related topic, then the probability of finding banana in that document is approximately 80%. This is not an exact probability measure since LDA models the relationship of a document to each of the different topics with different probabilities. In other words, a given document or message may be 75% related to the fruit topic and 25% related to the book topic.

To analyse the Enron corpus and find the distribution of topics per email message, the MACHine Learning for Language Toolkit (MALLET) was used. MALLET is open source software for machine learning which implements LDA. MALLET provides the necessary components to develop a pipeline of possible procedures when analysing data. Prior to running LDA on the Enron messages, a number of pre-processing steps were applied to the texts, including tokenization, lowercasing of tokens, lemmatization (using the Stanford Core NLP tool), removal of stop-words using a standard stop-word list, and removal of non-alphabetic tokens (since numbers may not be informative in the context of topic modelling). The email contents were also preprocessed to remove some irrelevant text. Specifically, where an email was forwarded, the embedded header information included in the email body was stripped.

To topic model the Enron emails, LDA requires the number of target topics as its input. We set the total number of topics to 20. We used MALLET, an open source software for topic modelling, which implements MALLET. MALLET provides the necessary components to develop a pipeline of possible procedures when analysing data. Prior to running LDA on the Enron messages, a number of pre-processing steps were applied to the texts, including tokenization, lowercasing of tokens, lemmatization (using the Stanford Core NLP tool), removal of stop-words using a standard stop-word list, and removal of non-alphabetic tokens (since numbers may not be informative in the context of topic modelling). The email contents were also preprocessed to remove some irrelevant text. Specifically, where an email was forwarded, the embedded header information included in the email body was stripped.

To topic model the Enron emails, LDA requires the number of target topics as its input. We set the total number of topics to 20, to use as an initial target set for the visualization, without careful investigation of the best number of topics. Table 1 provides an example of a topic (topic number 19) that MALLET has found in the Enron corpus, with the list of top 10 terms that have the highest
Figure 5. Shows the email communication between Enron users, zoomed into the year 2000. The analyst can see at a glance that there was substantial email traffic around Topic 19.

Figure 6. Shows the email communication between Enron users, zoomed into the year 2000 and topic 19. The topic details (i.e., a specification of key terms for the topic) are displayed when an analyst hovers over a topic label.

probabilities in the topic. From this list, one may intuitively decide that this topic is concerned with energy pricing. The alpha measure and total number of tokens are the probability of the topic (summing to 1.0 across all 20 topics) and the total number of tokens in the corpus that are most related to the topic, respectively.

After determining the topics, we map each email to its most probable topic by counting the proportion of words in the email associated with each topic.

**Assumptions in the Enron visualization**

We have applied some strategies to simplify the Enron data for the purpose of experimenting with the architecture of the
visualization approach. First, in terms of the users (individual people represented as senders or recipients in the dataset), we have made use of a “suspect list” developed to focus analysis on a specific subset of individuals. Each individual in the suspect list is represented as an individual user in the visualization. Other Enron-internal users (i.e., users with an email address @enron.com) are grouped together into a single node (Node 1). All external users (non-Enron email addresses) are also grouped together, into a single node (Node 0).

Second, where an email has multiple recipients (e.g., on a to: or cc: list), we treat that email as multiple (identical) individual messages connecting the sender and each recipient.

Implementation

CommViz is built using the Data-Driven Documents (D³) software package (D3.js). D3.js is a JavaScript library that gives users the capabilities to manipulate HTML documents using Scalable Vector Graphics (SVG) and Cascading Style Sheets (CSS). The resulting visualization is interactive, flexible, and accessible via any web browser.

We store the data, consisting of the meta-data of each email, its content, and the topic distribution for the email, in a MySQL database. The data is additionally pre-processed into tables corresponding to the core structures for the visualization – specification of the segments on the axes, the organization of nodes within the segments, and tables representing contexts, users, and timeframes. This database is accessed by a RESTful web service which queries the database and translates it into JSON format for D³. Two types of JSON object are produced, node objects (representing people/users) and edge objects (representing email communications between people). The web service is responsible for grouping the nodes by context and timeframe for display on the appropriate axes, and linking the relevant senders (source node) and recipients (target node) based on the header and context data for each message.

The JSON objects are structured as follows:

```json
{ // Node
  "nodeId": "100",
  "userId": "101",
  "email": "angela.cadena@enron.com",
  "name": "Angela Cadena",
  "type": "1" // axis
  "group": "0", // segment on axis
  "index": 2 // node index within segment
}

{ // Edge
  "source": {nodeId:100},
  "target": {nodeId:103},
  "numEmails": "64"
}
```

The type feature in the Node object represents the axis the node will appear on. A type value of 1 represents the primary user-context axis, 2 represents the cloned user-context axis, 3 the primary user-timeframe axis and 4 the cloned user-timeframe axis. Edges go from a sender source to a recipient target. Senders are drawn in axes 1 and 4 and recipients in axes 2 and 3. The web service groups related nodes by topic or timeframe (indicated with the group feature), depending which axis they belong to. Nodes are ordered within a segment group according to the index value. An Edge object specifies a source, a target, and the number of emails that are aggregated in the edge. The information about the sender of the message and the relevant context (group ID on axis 1 or 2) is in the corresponding source node object, and the recipient information and the relevant time (group ID on axis 3 or 4) is in the target node object, hence it is not necessary to capture them explicitly here; context and time can be inferred from group ID mappings.

In this way, the generation of the input objects for D³ is handled exclusively by the web service; leaving just the drawing process itself to the web application. The hive plot application draws the dots along the axes using the node data and generates the arcs (links) using the edge data.

Semantic Parallel Coordinates

We also experiment with an adaptation of Parallel Coordinates for multi-dimensional data, inspired by our experience with semantic hive plots. This adaptation also integrates two attributes along each axis, such that individual links capture four dimensions of the data. The result, using a variant of the D³ code developed for semantic hive plots, can be seen in Figure 7. This visualization differs from the semantic hive plots visualization only in the orientation of the axes (vertical cf. radial); the strategy of incorporating multiple dimensions in a single axis remains the same.

This adaptation can be seen as a generalization of the basic parallel coordinates approach to semantic coordinates. The links corresponding to the quadruple of {sender, recipient, time, context} appear in the center of the visualization. As for semantic hive plots, this offers the ability to observe patterns involving 3 (or 4) variables simultaneously. For instance, the number of messages sent between a given sender and a given recipient on a given topic can be compared to the number of messages sent between those same people on different topics. Multi-dimensional relationships such as these can only be represented in standard parallel coordinates through a poly-line, which are difficult to compare visually as there is no aggregation of dimensions.

It may further offer advantage over the radial layout of semantic hive plots, in that the distances between axes are
consistent (see Discussion section Visual scalability. However, radial layouts may be more effective when conveying quantitative information, i.e. when ordering segments along the axis in terms of quantitative characteristics, as has been argued with respect to hive plots. The intuition underlying this is that the length of a link connecting nodes on connected axes in a radial plot will convey visual information about the quantitative value of the link, which could aid in interpreting those quantitative values. We leave these issues as open questions for future research.

Enron analytical use cases

In this section, we introduce specific analytic use cases that are supported by CommViz, as seen from the example of the Enron data set visualization. The capabilities demonstrated here would generalize to any data set containing the same structural characteristics of sender, recipient, time, and topic, such as social media messaging or web discussion forums; we will provide further examples below.

As we have indicated, our primary focus question for the Enron dataset is “Who sent a message to Whom about what Topic in a given timeframe?”, or from a more aggregated perspective, “What topics are people talking about and when?”. The visualization is designed specifically to provide insight into this question, by organizing messages into topics (on the user-topic axis), and by connecting those topics to the times of those messages (links between the user-topic and user-timeframe axes).

The organization of information in CommViz thus presents contexts (topics) as the main organizing dimension along one axis, and time as the main organizing dimension along the other. Users are secondary, with a node for each user being duplicated within each segment along each axis. From this it follows that patterns relating to contexts (topics) and to time, and their interaction, will be most salient in the visualization.

Analysis of thematic patterns

There are several questions that follow from the main focus question that are also revealed by the visualization.

- **Which users had heavy communication on a given topic?** This corresponds to the links connecting the user-topic axis and its clone. A dark link connecting two users within a topic, between those axes, indicates a substantial number of emails between the two users on that topic.

- **What topic is associated with the largest number of messages?** This can be observed via the density and darkness of links originating in a segment along the user-topic axis.

- **During which time periods was a topic most prevalent?** This can be observed by looking at the links originating in the user-topic axis and ending in the user-timeframe axis. Again, density and darkness of links in that region of the coordinate space indicate important activity.

The overview visualization of the Enron dataset in Figure 3 reveals a number of characteristics at a glance:

1. **Topics 0 and 19 appear to be the most active topics, across a broad set of users.** This can be observed through the density of links originating in those topic segments.

2. **There appears to be one main user who is the target recipient of emails related to topic 19.** This is indicated by the darkness of color of the links leading into user 0. User 0 is the node for external individuals, hence this indicates that many emails related to topic 19 are going to recipients external to the company.

Analysis of temporal patterns

Viewed from another perspective, the layout also provides insight into temporal patterns in the data.

- **When has there been the most message activity?** This can be observed via the density and darkness of links entering a segment along the user-timeframe axis.

- **How many messages were sent between two users in a given timeframe?** This corresponds to the links connecting the user-timeframe axis and its clone. A dark link connecting two users within a time segment, between those axes, indicates a substantial number of emails between the two users at that time.

Considering the overview of the Enron dataset in Figure 3 again, we can see:

1. **There is likely a time anomaly in the underlying data.** There are messages corresponding to 1979, and then a jump to 1998. This is easily perceived in the distribution of links coming into the primary user-timeframe axis. It can also be observed through the cloned user-timeframe axis that there are only a small
Figure 8. Shows the email communication between Enron users, zoomed into topic 19. The light green segment corresponds to 2001 and the red to 2002.

Figure 9. The Enron network, displayed in a typical node/link force directed layout in the Cytoscape tool. Here, nodes are users and links are emails exchanged between two users connected by the link. Hence the two dimensions \{sender, recipient\} are captured in the nodes (without directionality). The topic of the message is captured through link color. If there are multiple topics connecting the sender and recipient, only the last link (topic) plotted will be visible. This is a limitation of the node/link visualization.

2. Emails relating to Topic 19 are fairly consistent across time, with an apparent increase in 2001, tapering off again in 2002. The density of links related to this topic increases in 2001; this is particularly apparent when we zoom into topic 19 as shown in Figure 8.

We contrast the ability to discern such patterns using CommViz with the node-link network in Figure 9. In this visualization, nodes correspond to users and links to emails between two users, although no directionality (i.e., the roles of sender/recipient) is captured. Topics are incorporated through link color. If there are multiple topics connecting the sender and recipient, only the last link (topic) plotted will be visible. The loss of information in this case is a limitation of overplotting that occurs in the node/link visualization. No temporal information is captured in this example; it is difficult to imagine how that could be incorporated in addition to the existing information.

**Extension to other datasets**

To demonstrate the flexibility of CommViz to accommodate the representation of other communication data sets, we applied it to a small Twitter sample, and to the “Dinofun World” visitor communication data\(^4\) from the VAST 2015 Mini-Challenge 2.\(^5\) Both adaptations can be viewed at blinded.\(^*\)

For the Twitter data set demonstration, we defined context as the source application used to post the tweet (e.g., Twitter for Android, or TweetDeck). The data mapped straightforwardly into CommViz node and link structures; we note that by defining context as was done here we have ignored the thematic content of the tweets themselves, which could be viewed as a fifth dimension in this representation. We have not considered how best to incorporate an additional dimension of the data.

For the VAST data set, context represents the location of the sender of a message, following the structure of the data (originator=sender ID, recipient ID, timestamp, park area from which message was sent). It includes 28,074 links representing 9,616 messages between 2,074 people and again mapped straightforwardly into the CommViz structures. However, the size of this data set resulted in a very dense visualisation as shown in Figure 10; strategies for pruning the dataset (e.g., based on simple thresholds of minimum number of messages sent) would likely be required. In Figure 11, we can see the value of zooming for exploring specific subsets of the data; the nodes spread out along the axes, and links separate.

**User study**

To assess the value of CommViz to support visual analytics, we conducted a user study. We compared CommViz directly to the force-directed layout commonly used for communication networks, as illustrated in Figure 9 above.

**Evaluation approach**

Evaluating the effectiveness of visualizations using controlled experiments to compare different techniques/tools\(^51\)–\(^54\) or the same technique with different settings or styles\(^55\) is the most common type of visualization evaluation. However, very few techniques or tools are built for identical purposes in the same domains, and on similar or the exact same types of data.\(^56\)

While most traditional user-centered studies were carried out in collaboration with known “expert” users in the domain, recently crowdsourcing has been explored for such evaluations,\(^57,58\) resulting in a greatly increased and diversified set of users.\(^59\) We also chose to make use of a crowdsourcing methodology, utilizing Amazon’s Mechanical Turk Service.\(^57\)

\(^*\) Source data set is selectable through the drop-down menu at top left.
Verspoor, Ologhi, Robles Granda

Figure 10. Shows the application of CommViz to the VAST 2015 Mini-Challenge 2. Here contexts correspond to different locations in an amusement park. Context 5 is “Wet Land”. Adapting CommViz to this data was straightforward using a “middleware” layer supporting mapping from a data file to the elements of the visualization. This data set is large; over 28000 links are displayed in this screen shot. This example highlights the need for further support for filtering links based on the number of messages.

The type of visualization evaluation we adopted for CommViz is an analysis-oriented task where two different visualizations of the same data are compared with each other with respect to their effectiveness in supporting users to correctly answer data-driven questions. This is in contrast to studies which ask users to interact with the visualization and rate it based on a collection of pre-defined criteria; the assessment is oriented to the utility of the tool for visual analytics, as a step in an analytical workflow, rather than capturing user preferences. Since the force-directed layout (cf. Figure 9) has remained among the most frequently used methodologies for visualizing communication networks, the effectiveness of CommViz was compared with that of this rival visualization methodology as implemented in Cytoscape.

We first captured six screenshots from each visualization (CommViz and Cytoscape) related to the Enron email corpus. Each screenshot provided a view of the same subset of the data, required to answer a particular type of question; the main focus of each question is listed in Table 2. Figure 3 is used for the CommViz image for Question 1, while Figure 9 provides an example of the Cytoscape representation for Question 1. All questions/images are detailed in Supplementary Files 1 and 2. The questions cover a variety of topical and date information in the data set. The screenshots were accompanied by a description to help the user understand the context of the question, as well as the multiple choice question itself. The Mechanical Turk Worker was asked to select an answer from one of four options, with the correct answer known based on standard analytical querying of the data. Users answered the 12 questions, in one of three different randomly determined orders, shown in Table 3.

Finally, the results obtained for the entire set of answers for all three sequences of questions were statistically analysed to compare the effectiveness of CommViz with that of Cytoscape. While overviewing, dynamic filtering, and data-aware zooming are among features that can be utilized as metrics for technique-centered visualization comparisons, we made use of accuracy and inter-rater agreement measures as evaluation metrics, due to our emphasis on the analytical utility of the tool.

Mechanical Turk tasks and settings

Each of the three sequences of 12 questions was put together with the corresponding descriptions and screenshots into a single Human Intelligence Task (thus there were three HITs in total). The questions in each HIT appeared on the screen one at a time (with the question number, description, visualization screenshot, the question, and the four possible answer options) and the users were not able to return to a
Figure 11. Shows the application of CommViz to the VAST 2015 Mini-Challenge 2. In this example, zooming into one context (5) and a specific day (6/6/2014) results in the spread of users along the User-Context dimension.

Table 2. The main focus of each question created on the basis of the Enron email corpus

<table>
<thead>
<tr>
<th>ID</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Topic overview: to identify the most prominent topic in the dataset</td>
</tr>
<tr>
<td>2</td>
<td>Year overview: to identify the year with the smallest number of emails</td>
</tr>
<tr>
<td>3</td>
<td>Emails of Topic 19: what % in 2001</td>
</tr>
<tr>
<td>4</td>
<td>Emails of the year 2001: to find the topic with the second largest number of emails</td>
</tr>
<tr>
<td>5</td>
<td>Emails of Topic 19 and the year 2001: to identify the distribution of emails in different months</td>
</tr>
<tr>
<td>6</td>
<td>Topic14 vs. Topic17: to compare the distribution of emails in different years</td>
</tr>
</tbody>
</table>

Table 3. The three random display sequences for the 12 data-driven questions. Each $A_i \ (i \in \{1, \ldots, 6\})$ shows a question accompanied with a Cytoscape visualization and each $B_j \ (j \in \{1, \ldots, 6\})$ represents a question with a CommViz screenshot. The numbers match with the question numbers in Table 2.

<table>
<thead>
<tr>
<th>Sequence #</th>
<th>q1</th>
<th>q2</th>
<th>q3</th>
<th>q4</th>
<th>q5</th>
<th>q6</th>
<th>q7</th>
<th>q8</th>
<th>q9</th>
<th>q10</th>
<th>q11</th>
<th>q12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B5</td>
<td>B3</td>
<td>A6</td>
<td>B4</td>
<td>A1</td>
<td>A4</td>
<td>B2</td>
<td>A5</td>
<td>B1</td>
<td>A6</td>
<td>B3</td>
<td>A2</td>
</tr>
<tr>
<td>2</td>
<td>B3</td>
<td>A1</td>
<td>A4</td>
<td>B6</td>
<td>A3</td>
<td>A5</td>
<td>B2</td>
<td>B5</td>
<td>A6</td>
<td>B1</td>
<td>A2</td>
<td>B4</td>
</tr>
<tr>
<td>3</td>
<td>A6</td>
<td>B3</td>
<td>A5</td>
<td>A4</td>
<td>B5</td>
<td>B6</td>
<td>B4</td>
<td>B1</td>
<td>A2</td>
<td>A3</td>
<td>B2</td>
<td>A1</td>
</tr>
</tbody>
</table>

previous question once they had answered it and moved to the next.

Each HIT was submitted to Mechanical Turk separately to be answered by 50 distinct Mechanical Turk workers. We implemented several qualification criteria for the workers including:

- The workers were not allowed to take any of the other HITs if they had already answered the questions of another HIT.
- The worker had to possess a minimum of 10,000 approved HITs.
- The worker had to have a minimum of 98% approval rate on the HITs they had previously completed for any requester on Mechanical Turk.
We used the last two qualification criteria to obtain experienced Turkers, rather than limiting the set of qualified workers to those with "expert" standings which would slow down the process of getting the HITs completed due to the smaller pool of experts. We awarded $1 for each completed assignment per worker. The HITs were set to expire in 7 days after submission. All three HITs were completed in the given time period. We approved all 150 assignments performed by the workers.

**Evaluation results and discussion**

Once the results of the completed HITs were received from Mechanical Turk workers, the answers were analysed to understand the effectiveness of CommViz in enabling users to identify the correct answers for the six questions compares with that of Cytoscape.

The first metric for effectiveness analysis of the visualization techniques was accuracy. The accuracy per worker per visualization was calculated as the proportion of the six responses based on the visualization that were correct, with respect to the gold standard answers. The overall accuracy of each visualization technique was calculated as the average of these values across the 50 individual workers participating in a given HIT.

The same technique was applied to calculate the overall accuracy over the 3 HITs (i.e., 150 sets of answers). To measure agreement of different workers, Fleiss’ Kappa was calculated between the answer sets for each visualization technique (both within a HIT and across all 3 HITs). To find whether the differences between accuracy values of the two visualizations were statistically significant, we carried out paired t-tests between the accuracy values related to each HIT and overall. For this, the accuracies of the two visualizations for the six questions were used in a similar procedure to the resampled paired t-test of Dietterich et al (1998). In our case, a question was considered as a test portion in a trial. The question-level accuracies were used as observed proportions of samples correctly answered by Mechanical Turk workers.

Table 4 summarizes the accuracy and inter-rater agreement of each technique for each HIT and overall. As shown in these results, in all three HITs, as well as when considering all of the answers together, CommViz outperforms Cytoscape with respect to accuracy. These differences in accuracy are statistically significant at the standard 95% confidence level, based on a paired t-test.

However, except for the first HIT, in other cases, workers seem to have agreed with each other more in the case of Cytoscape, as indicated with higher Fleiss’ Kappa values. These inter-rater agreement values do not consider the gold standard answers and only consider the agreement among the distinct workers when answering the data-driven questions for each visualization technique. This can be interpreted to mean that there was more consistency in responses among users with Cytoscape visualizations, but that these responses were generally not correct, i.e., that the users were misled by the visualization in a consistent way.

To understand the average agreement rate of the answers given by the workers with respect to the gold standard answers, we further analysed the results in two other ways. First, the average Fleiss’ Kappa was calculated for each visualization, and for each question, as compared with the gold standard answers. This focuses on the answers given for each individual question per visualization. In this case, there are six agreement values, one for each question as compared to the gold standard answer of the same question (i.e., two subjects and 50 responses for each HIT). The average of these six agreement values was then calculated for the HIT.

Second, the average Fleiss’ Kappa was calculated for each visualization, and for each worker, as compared to the gold standard answers. This kappa value puts emphasis on the average agreement of the users’ six answers with the six gold standard answers for the six questions that were accompanied by screenshots of each visualization technique. For this, 50 agreement values between two sets of six answers were calculated (i.e., six subjects and two raters). The average of these 50 agreement values was then calculated for each HIT. The results of these two agreement analyses are given in Table 5.

The results in Table 5 suggest that CommViz provides higher agreement rates when the worker answers are compared against the gold standard answers. Although the agreement rates are very low for both techniques when focusing on individual questions (as indicated by negative Fleiss’ Kappa values), still the agreement values are higher for CommViz. When considering individual raters’ answers against our gold standard answers, the agreement rates by both techniques are higher and in addition, those of CommViz are all in the range of fair agreement.

We further analysed the results at the individual question level to identify where each of the two visualization techniques shows superior performance; in other words, what type of questions can more effectively be answered by users using one visualization technique as compared to the other. The accuracy of each visualization was first calculated for answers given by the Mechanical Turk workers for each question. Similar to the technique-level analysis, the gold standard answers for each question were used to calculate accuracy. Table 6 summarizes the results for each visualization technique, per HIT, per question. The chi-square values were calculated between the categorical answers given per question for each visualization technique to measure the statistical significance of the differences.

As shown in Table 6, CommViz consistently performs better for Questions 2, 4, 5, and 6, except in the second HIT. For Questions 4 and 5 in the three HITs, the differences between higher performances of CommViz and lower accuracies of Cytoscape are statistically significant. As a reminder, the reader may refer to Table 2 for descriptions of the questions, and to the Supplementary Files 1 and 2 for the actual questions with screenshots.

Overall, CommViz performs better than Cytoscape and for Questions 4, 5, and 6, the differences are statistically significant, although for the other two questions (i.e., 1 and 3), Cytoscape outperforms CommViz. Another analysis was carried out to understand whether the workers were confused by similar answer options for some of the questions when perceiving each of the two visualizations. For this, the raw distribution of total worker answers per question was analysed as shown in Figure 12(a) and Figure 12(b) for CommViz and Cytoscape, in the left and right panel, respectively.
Table 4. The accuracy and inter-rater agreement analysis of the results achieved using Cytoscape and CommViz. A † sign in front of each accuracy value indicates the statistically significant difference between that value and the accuracy of the competing visualization based on a paired t-test at the standard 95% confidence level. In all HITs and overall, CommViz users achieve statistically significant higher accuracy on the test questions than Cytoscape users.

<table>
<thead>
<tr>
<th>HIT #</th>
<th>#Approved answers</th>
<th>Visualization</th>
<th>Avg. accuracy</th>
<th>IRA: Fleiss' Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>Cytoscape</td>
<td>0.4667</td>
<td>0.2131</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>CommViz</td>
<td><strong>0.6033†</strong></td>
<td>0.2176</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>Cytoscape</td>
<td>0.4667</td>
<td>0.2116</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>CommViz</td>
<td><strong>0.5233†</strong></td>
<td>0.1422</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>Cytoscape</td>
<td>0.4267</td>
<td>0.2004</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>CommViz</td>
<td><strong>0.5333†</strong></td>
<td>0.1605</td>
</tr>
<tr>
<td>Overall</td>
<td>150</td>
<td>Cytoscape</td>
<td>0.4533</td>
<td>0.2114</td>
</tr>
<tr>
<td>Overall</td>
<td>150</td>
<td>CommViz</td>
<td><strong>0.5533†</strong></td>
<td>0.1743</td>
</tr>
</tbody>
</table>

Table 5. Fleiss' Kappa agreement analysis of individual questions and individual raters versus gold standard answers for each visualization technique.

<table>
<thead>
<tr>
<th>HIT#</th>
<th>Question vs. gold standard</th>
<th>Rater vs. gold standard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cytoscape</td>
<td>CommViz</td>
</tr>
<tr>
<td>1</td>
<td>-0.2565</td>
<td>-0.1680</td>
</tr>
<tr>
<td>2</td>
<td>-0.2565</td>
<td>-0.2007</td>
</tr>
<tr>
<td>3</td>
<td>-0.2782</td>
<td>-0.1987</td>
</tr>
<tr>
<td>Overall</td>
<td>-0.2604</td>
<td>-0.1844</td>
</tr>
</tbody>
</table>

Table 6. Question-level accuracy analysis of the answers given by Mechanical Turk workers for the two visualization techniques. Question numbers match with those in Table 2. In cases where CommViz has outperformed Cytoscape, the question number has been highlighted in bold face. Asterisks show statistically significant Chi square values at the standard 5% confidence level where the degree of freedom is \((2-1)(4-1) = 3\) for 2 visualization techniques and 4 possible answer options per question.

<table>
<thead>
<tr>
<th>Question</th>
<th>HIT#</th>
<th>Visualization</th>
<th>Chi square</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cytoscape</td>
<td>CommViz</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.8400</td>
<td>0.8000</td>
</tr>
<tr>
<td>2</td>
<td>0.8200</td>
<td>0.6800</td>
<td>4.5700</td>
</tr>
<tr>
<td>3</td>
<td>0.8800</td>
<td>0.7000</td>
<td>5.746</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.4400</td>
<td>0.4400</td>
</tr>
<tr>
<td>2</td>
<td>0.5400</td>
<td>0.4400</td>
<td>5.1252</td>
</tr>
<tr>
<td>3</td>
<td>0.3800</td>
<td>0.6200</td>
<td>7.5256</td>
</tr>
<tr>
<td>Overall</td>
<td>1</td>
<td>0.6000</td>
<td>0.4800</td>
</tr>
<tr>
<td>2</td>
<td>0.6200</td>
<td>0.4000</td>
<td>7.6202</td>
</tr>
<tr>
<td>3</td>
<td>0.5400</td>
<td>0.4000</td>
<td>4.5993</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.0600</td>
<td><strong>0.5200</strong></td>
</tr>
<tr>
<td>2</td>
<td>0.0800</td>
<td>0.5200</td>
<td>25.2848*</td>
</tr>
<tr>
<td>3</td>
<td>0.0800</td>
<td>0.3800</td>
<td>23.398*</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.4600</td>
<td><strong>0.6600</strong></td>
</tr>
<tr>
<td>2</td>
<td>0.4400</td>
<td>0.6400</td>
<td>10.2761*</td>
</tr>
<tr>
<td>3</td>
<td>0.4200</td>
<td>0.6400</td>
<td>15.0925*</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0.4000</td>
<td><strong>0.5200</strong></td>
</tr>
<tr>
<td>2</td>
<td>0.3000</td>
<td>0.4600</td>
<td>4.3935</td>
</tr>
<tr>
<td>3</td>
<td>0.2600</td>
<td>0.4600</td>
<td>4.5111</td>
</tr>
<tr>
<td>Overall</td>
<td>1</td>
<td><strong>0.8467</strong></td>
<td>0.7267</td>
</tr>
<tr>
<td>2</td>
<td><strong>0.4533</strong></td>
<td>0.5667</td>
<td>6.0350</td>
</tr>
<tr>
<td>3</td>
<td><strong>0.5867</strong></td>
<td>0.4267</td>
<td>15.5261*</td>
</tr>
<tr>
<td>4</td>
<td>0.0733</td>
<td><strong>0.4733</strong></td>
<td>85.5994*</td>
</tr>
<tr>
<td>5</td>
<td>0.123</td>
<td>0.4400</td>
<td><strong>0.6467</strong></td>
</tr>
<tr>
<td>6</td>
<td>0.123</td>
<td>0.3200</td>
<td><strong>0.4800</strong></td>
</tr>
</tbody>
</table>

As depicted in Figure 12(b), the visualization in Cytoscape is sufficient for the workers to select among competing answer options for Questions 1 (topic overview) and 3 (% of emails in a given topic corresponding to a given year) where this technique outperforms CommViz. However, for the other questions, the force-directed graph layout in Cytoscape did not support the users to fully grasp the information in the visualizations, as indicated by more uniform distributions of answers by workers.

The answer distributions in Figure 12(a) shows that for Questions 4, 5, and 6 (each involving analysis of distributions over different combinations of attributes), where CommViz has significantly outperformed Cytoscape, there are overall decisive selections by workers; there are no strongly
competing answer options for the correctly selected options. For Question 3, the second competing option is much closer to the first and correct answer option. For the other question (Question 1) where CommViz was outperformed by Cytoscape, however, there is a large gap between the number of workers that selected the correct option and the number of workers that selected other options. Yet, the number of workers who chose one of the wrong options was greater than those who made the same mistake when looking at Cytoscape’s visualization of the same data. For Question 2 (year overview), CommViz also outperformed Cytoscape; CommViz enabled users to identify the correct answer as Option 2 whereas Cytoscape has confused the users between Options 1 and 2.

In general, our results demonstrate that CommViz enables users to have a more effective understanding of the information content of the underlying data. At the same time, users appear to be more certain in their answer when using the CommViz visualizations; there is more consistency in the responses across users than with the Cytoscape-based visualization. We have shown that CommViz has important advantages over the standard force-directed layout communication visualization, in particular with question types that involve comparative analysis between contexts/topics or across time.

Discussion

Visual scalability

The requirements of visual analysis of large-scale data, in particular with the complexity of 3-4 dimensions of relevant information, pose substantial challenges to visualization strategies. We have emphasised in our approach analysis of trends that involve the multivariate nature of the data and designed a visualization approach that specifically accommodates integrative analysis. However, in so doing, we have introduced some limitations that impact the visual interpretability of some of the data. Specifically, the space in between the axes connected by the links grows the further you move from the axis origin; hence there is an inherent difference in interpretability in the links originating from the first and last groups along the same axis. This problem is mitigated to some extent by the ability to zoom into the groups, to examine the relevant subset of the data, as the nodes within that group will spread out along the axis and links not relevant to that subset of the data will drop away (as illustrated in Figure 11).

Temporal dynamics

While topic modelling on its own provides a summary of the themes represented in the data set that can be visualized in various ways, and the number of messages corresponding to particular topics over time can be represented in a topic-centered visualization such as TIARA, the individual-level patterns relating to the topics (e.g., Who is talking about What topics?) cannot be easily interrogated without explicitly representing individuals in the representation. CommViz allows for this level of analysis, in addition to the static snapshot of general topic trends over time, in the links connecting the two primary axes.

Visually, contexts are organized somewhat arbitrarily in CommViz along the user-context axis. For topics, we could explore the incorporation of a visual notion of topic strength such as used in the TIARA system. Since the prevalence of a topic is expected to change over time, this would reveal another aspect of the dynamics of the communication network. Other organizations for different kinds of contexts, for instance to indicate proximity for the geographical locations, could be considered.

Inverted views and hierarchical organization

As described above, contexts and time are the primary organizing dimensions in the visualization. It is possible to imagine that it would be valuable to invert this, and “roll-up” the data in the opposite order; i.e., to map each user to a segment on an axis, and represent each context within that segment. This would enable investigation of user-oriented characteristics, such as contexts which are particularly prevalent for the given user, or timeframes in which they have been particularly active in communicating with another user.

We experimented with incorporating additional axes into the core visualization to capture this inverted organization; the context-user axis (and its clone) and the timeframe-user axis (and its clone). In principle, adding axes is straightforward to do in hive plots (and D3.js) and the generation of the appropriate node and edge definitions in JSON by the web service was a simple extension to
the existing methods. In practice, however, this resulted in a substantially increased number of nodes and links that needed to be rendered in the visualization, slowing the performance of the tool down. It furthermore resulted in a rather cluttered interface; the number of different data dimensions and connections among them became somewhat overwhelming.

An alternative approach may be to allow the analyst to select from various views of the underlying data, depending on the specific questions of interest. Then the primary and secondary organizing dimensions would be generated according to that selection; the web service would produce the nodes and edges and their groupings accordingly. The visualization is flexible enough to support multiple hierarchical groups along the different axes, and the web service approach to our implementation allows for these different views to be generated dynamically as needed.

One enhancement that builds further on the semantic characteristics of the data would be to group users into organizational structures, e.g. based on teams, divisions, etc. and according to user level, aggregating message traffic across, between, and within those groups. Such organizational structure has been shown to be valuable for deriving insights into organizational communications. The simplest integration of this into CommViz would be to capture this hierarchy in the representation of Users, and enable drilling down in a manner analogous to the Time dimension.

Future directions

CommViz is under active development, and there are a number of improvements that can be explored. Many aspects of the interface and visualization could be enhanced to provide more immediate feedback to users of the visualization.

One obviously needed refinement to the current application to Enron is to improve an analyst’s ability to understand what the topics along the user-topic axis correspond to. While we can provide a list of the words in the topic, which gives some intuitive sense of the topic, some other visualization of topics may help to convey their meaning more informatively. We may be able to take inspiration fromTermite or Themail for this.

Communication networks have many attributes that could be interesting to analyse visually; we have focused on the most salient questions and data for our initial investigation. However, we can imagine various modifications to the tool that make use of screen real estate in various different ways, to support different analyses. For instance, we have used the darkness of link color to convey density of communications between two axes. Perhaps some other link coloring scheme would be more appropriate, e.g., to convey information about the users connected by the link (either their network properties, e.g., centrality, or semantic properties such as whether the users belong to the same or different organizations). Other ideas that could be explored include adjusting what is displayed when an analyst hovers over an edge in the visualization. Currently, as seen in Figure 4, only the meta-data of the set of messages is displayed in the box at the top left of the screen. Another option would be to show the content of individual messages reflected by that link in some form, e.g., in some summary visualization such as a word cloud or as an entry point to a complete listing of the message content.

We have focused on zooming timeframes, to enable more detailed analysis of those timeframes, as the density of links at the scale of the full data set makes it difficult to interrogate individual relationships. Zooming of contexts is a straightforward extension, if the semantics of zooming into contexts can be clearly defined. It may also be important to enable investigation of specific users. The inverted view described above would support this straightforwardly.

The topic modelling algorithms used as the basis for organizing the messages thematically can be refined. While we make use of a simple Message-Topic model, a model that incorporates individuals in the model allows for interpretation of people in terms of the roles that they play in the context of the message collection, for instance which business area they are in, or whether their job is administrative or technical. McCallum proposed the Author-Recipient-Topic (ART) model for this and demonstrated improvements in classification tasks using this model. This model could be used to improve the topic modelling itself, and may be able to be exploited to organize user nodes within axis segments in a coherent way, i.e., in terms of their roles, rather than the current arbitrary ordering.

In general, incorporation of user interactivity to reflect the objectives of the user’s analytical questions, much as the capabilities available in PivotGraphs, Semantic Substrates, and DOSA, would make the tool more powerful. Attributes to aggregate, properties used to filter links, selection of the specific axes of interest and strategies for node placement could be specified by the user and applied dynamically. The use of structural characteristics such as node degree could be incorporated to organize nodes along the axes; while senders and recipients are currently ordered by arbitrary identifiers, we could order them by their social network characteristics. We look forward to exploring in more depth the aspects of CommViz that can be improved with deeper inspiration from these alternative approaches.

Finally, we plan to perform more user studies of the approach. As has been noted for parallel coordinates in particular, user-centred evaluations that address usability for different tasks are relatively rare but important. Our work has raised some obvious research questions that are suited to user studies, including a comparison of radial and parallel layouts for the semantic coordinate axes, of the impact of different choices for primary organizing dimensions on various tasks, and of various alternatives for placing entities or objects along the axes. While our current study employed naïve users from Mechanical Turk with realistic but synthetic analytical tasks, we plan to perform a situated study with professional data analysts studying real data and having specific analytical questions of interest.

Conclusions

This paper has described the CommViz communication network visualization system. We have introduced an approach to capturing several different dimensions of a complex communications dataset simultaneously within an accessible visualization, specifically the social network.
characteristics of the data (i.e., senders and receivers of messages), temporal information, as well as context, such as themes inferred from the content of the messages. Prior work has largely focused on only one of these dimensions, or in some cases two (time along with one of the other two dimensions), but to our knowledge this is the first work that integrates all four dimensions into one visually interpretable view of the data.

Our solution to representing this complex data has been achieved through adaptation of the hive plots visualization method, avoiding the lack of interpretability of typical node-link network visualizations. Hive plots provide an approach to organizing information that compactly captures a large number of data elements in a multi-dimensional coordinate system. In our semantic hive plots, we have mapped structural characteristics of individual messages to elements in the visualization, resulting in a visualization which includes immediately accessible fine-grained message-level information but enables identification of visual patterns by a human analyst across a large collection of messages.

We have applied CommViz to the Enron email corpus, briefly introduced its application to other message data sets, and demonstrated how it can support straightforward identification of meaningful patterns and relationships in that data. The CommViz approach will need to be tested with analysts with real questions about a real-world data set; this will surely result in design changes to the interface that will improve the analysts’ ability to notice patterns in the data visually, to observe greater temporal dynamics, and to interrogate the visualization flexibly. However, the progress we have made in developing a simple and interpretable visualization capturing the four semantic dimensions of the data, and that corresponds neatly to questions an analyst might naturally ask of the data, lays a strong foundation for future development.

Supplementary Files

We have provided several supplementary files, as listed.

1. CommViz screenshots used for the User Study. (PDF)
2. Cytoscape screenshots used for the User Study. (PDF)
3. Parallel Coordinate visualization of the Enron corpus. (PDF)
4. Raw and summary results of the User Study. (Excel)
5. README file for the program to populate a mysql database with the appropriate structures for use with CommViz, from a CSV file for a data set (file2DB-uniform.py,README.txt)
6. The Python code for the program to populate a mysql database with the appropriate structures for use with CommViz, from a CSV file for a data set (file2DB-uniform.py).

References


34. Gannett H. General summary showing the rank of states by ratios 1880, 1883. URL http://www.davidrumsey.com/luna/servlet/detail/RUMSEY-8-1-328903-1152181.


Acknowledgements
We gratefully acknowledge the work of The University of Melbourne students Bin Xu and Di Wang for their assistance, respectively, in setting up the user experiment on Mechanical Turk and extending CommViz to data sets beyond Enron.

Declaration of conflicting interests
The Authors declare that there is no conflict of interest.

Funding
This work was supported by the Bioterrorism Preparedness task of the Land Personnel Protection Branch, Land Division of the Australian Defence Science and Technology Group (DST Group).