NATIONAL ARCHIVES AND RECORDS ADMINISTRATION (NARA)
EMAIL NETWORK VISUALIZATION AND ANALYSIS PILOT
PROGRESS REPORT

June 11, 2005

SUMMARY OF DATA SELECTION

We investigated numerous possible sources of electronic message collections and determined that the Enron electronic mail collection made available by FERC (Federal Energy Regulatory Commission) was sufficiently similar in nature and style to the specified email collection to yield useful residuals. We then installed this dataset on our servers and imported it into our databases for analysis. We performed the same set of analysis tasks on the Enron dataset that we had done on the previous datasets. The result is a set of non-sensitive residuals in the form of visualizations that may be publicized to illustrate the capabilities of the project.

We looked into a number of available sources of online electronic message collections, including the visualization-related electronic mailing list archives that NCSA operates. The centralized communications model of these mailing lists (where message replies are broadcast to the entire group, instead of to the single author of the originating message) did not preserve enough of the social communications network of the email collection to yield good residuals in our analysis. It was eventually determined that the Enron email collection would provide the closest non-sensitive match to the specified collection.

The Enron email collection is a corpus of more than a half million email messages, representing around 92% of Enron staff emails released by FERC as part of its Western Energy Markets investigation. For the purposes of this project, we are using a copy of the dataset released by CMU in March 2004 that contains 517,431 emails from 151 users sorted into 5335 folders. There are 20,330 unique authors and 134,744 unique threads in the dataset, representing a strong mix of topics and communications styles very similar to the specified collection.

EXAMPLE VISUALIZATIONS

The dataset was downloaded and imported into our servers for analysis. The first analysis performed was tracing the occurrence of “California” (for the California energy crisis) through time in the collection. The following graphs illustrate various views of the resulting graph, demonstrating the wide range of visualization formats supported by the system and differing perceptions of each.

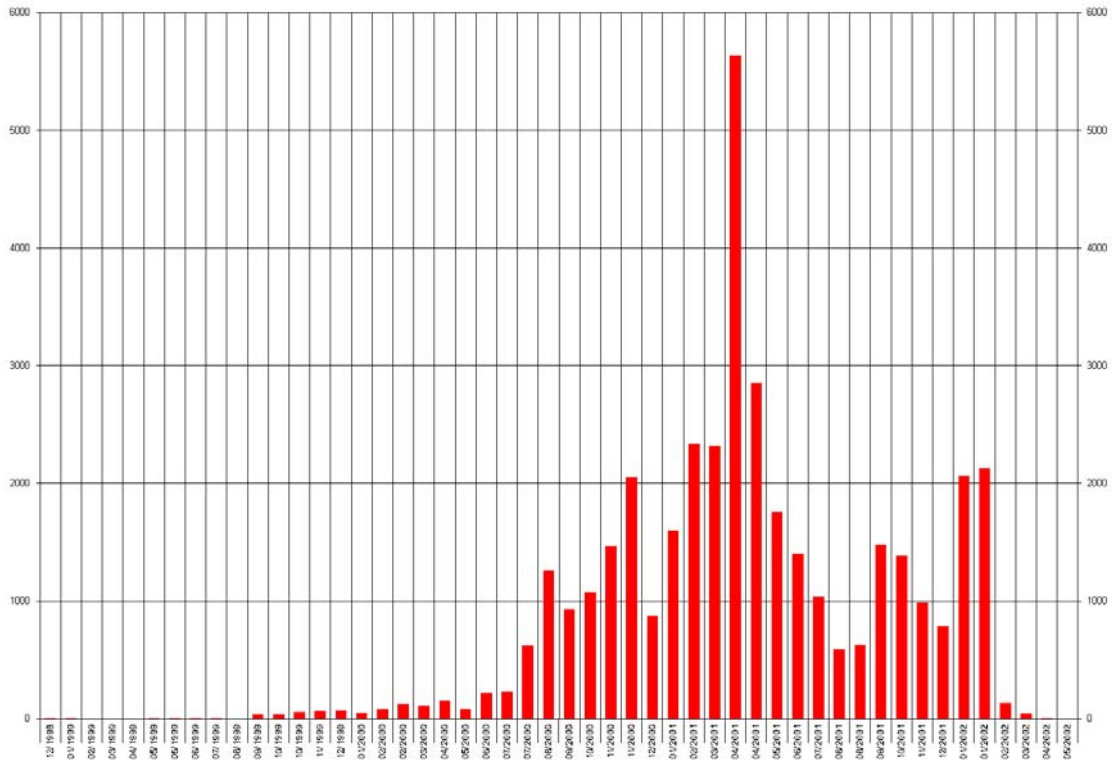

Figure 1 - Keyword tracing of "California" (2D Bargraph View)

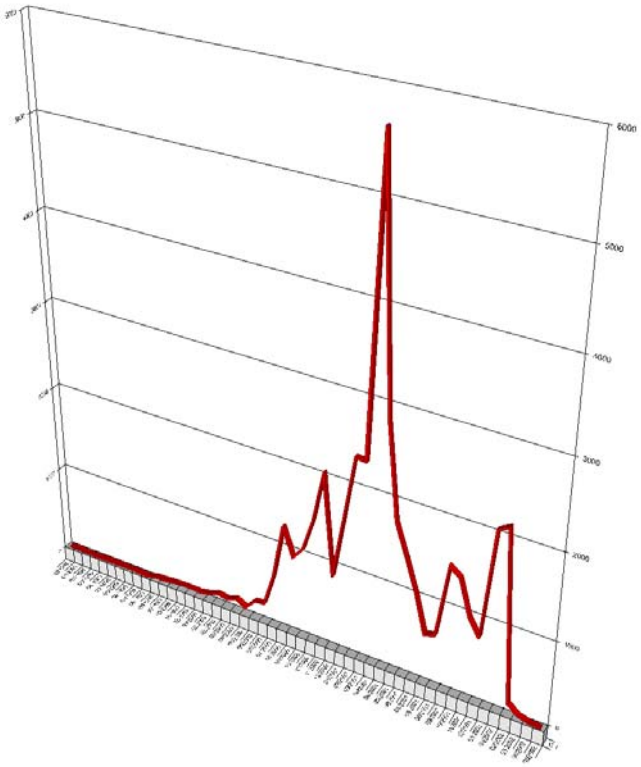

Figure 2 - Keyword tracing of "California" (3D Line View)

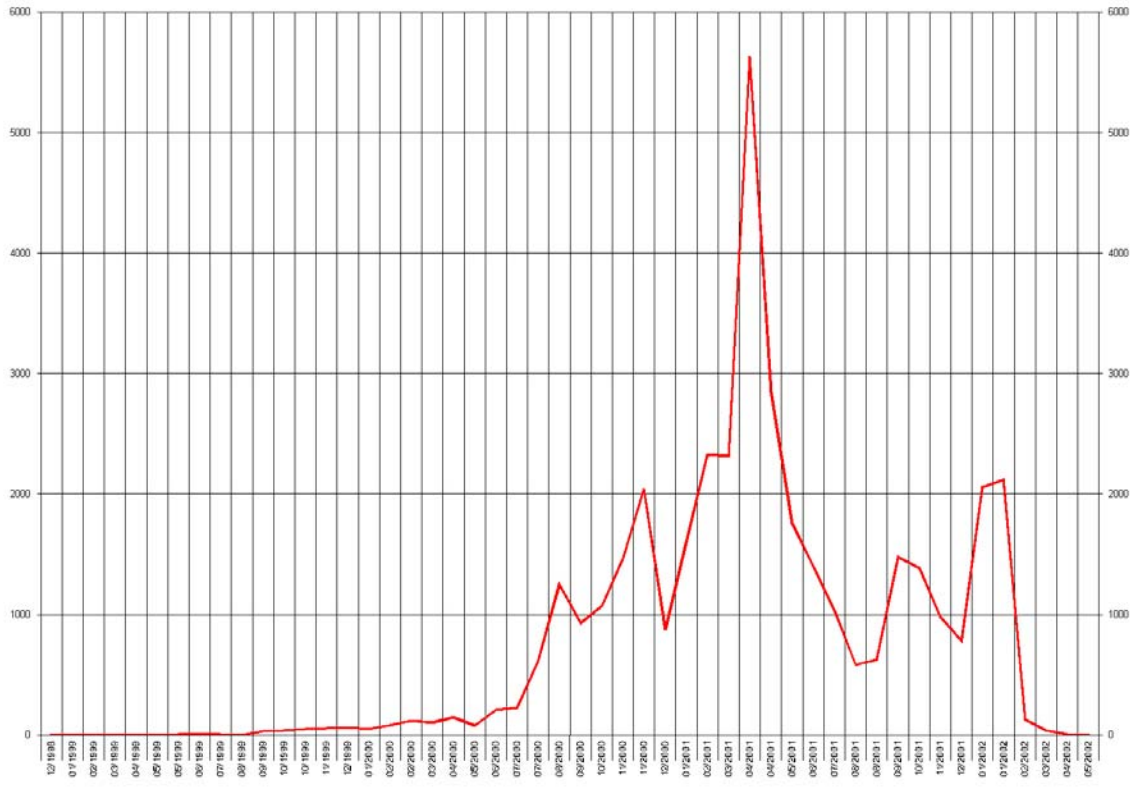

Figure 3 - Keyword tracing of "California" (2D Line View)

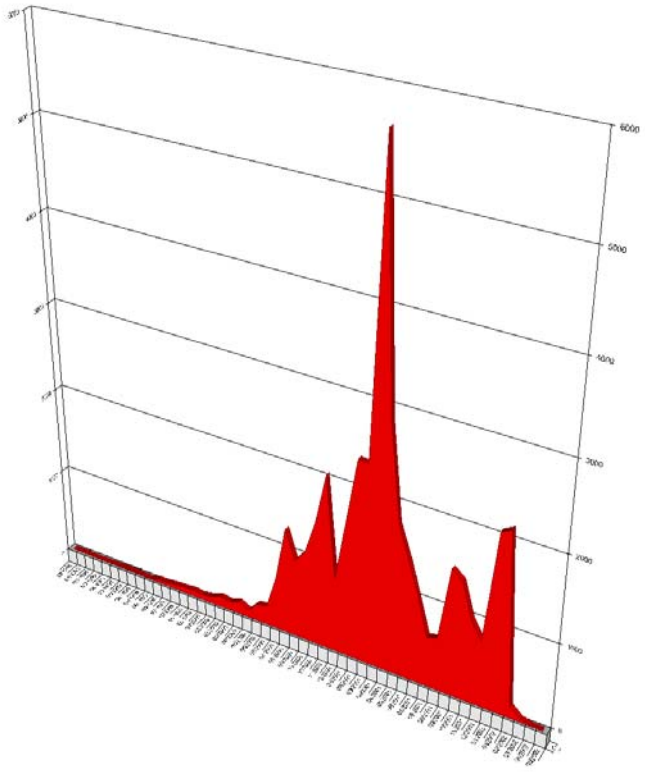

Figure 4 - Keyword tracing of "California" (3D Area View)

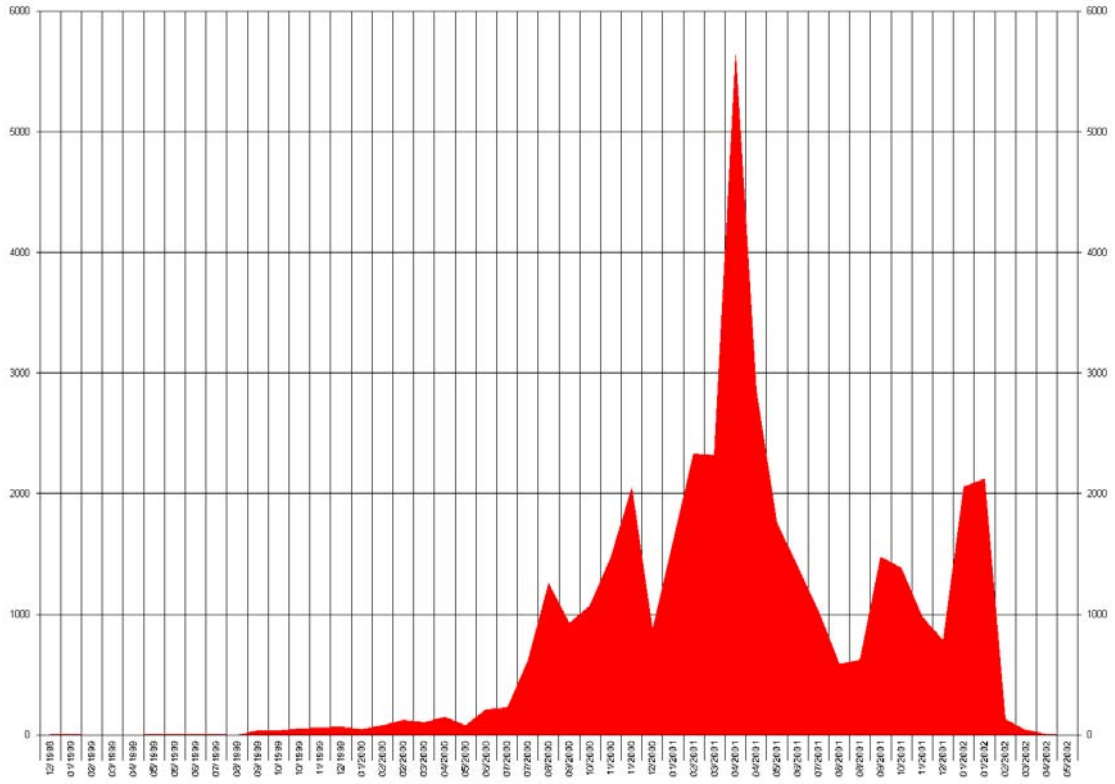

Figure 5 - Keyword tracing of "California" (2D Area View)

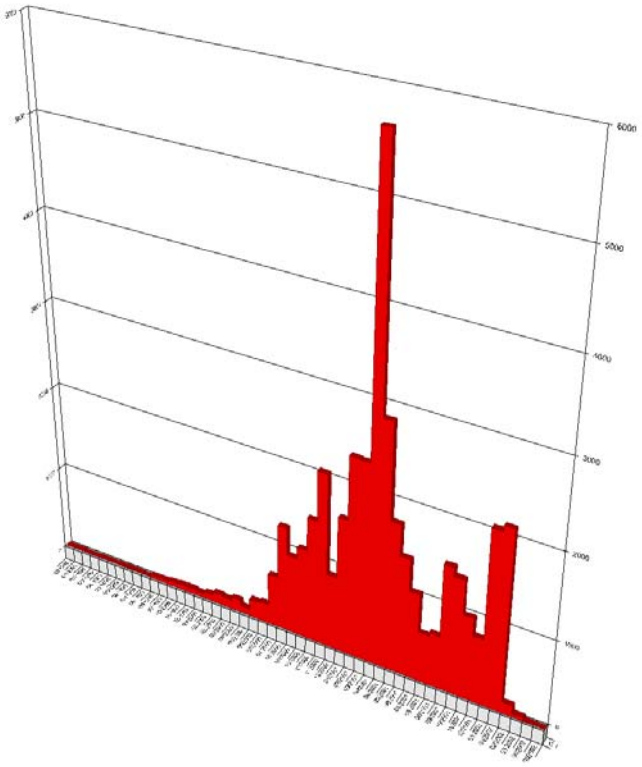

Figure 6 - Keyword tracing of "California" (3D Step View)

As a secondary test, the keyword “shred” was traced through the collection, illustrating fairly regular peaks with a steady increase in the number of occurrences between February 2000 and April 2002.

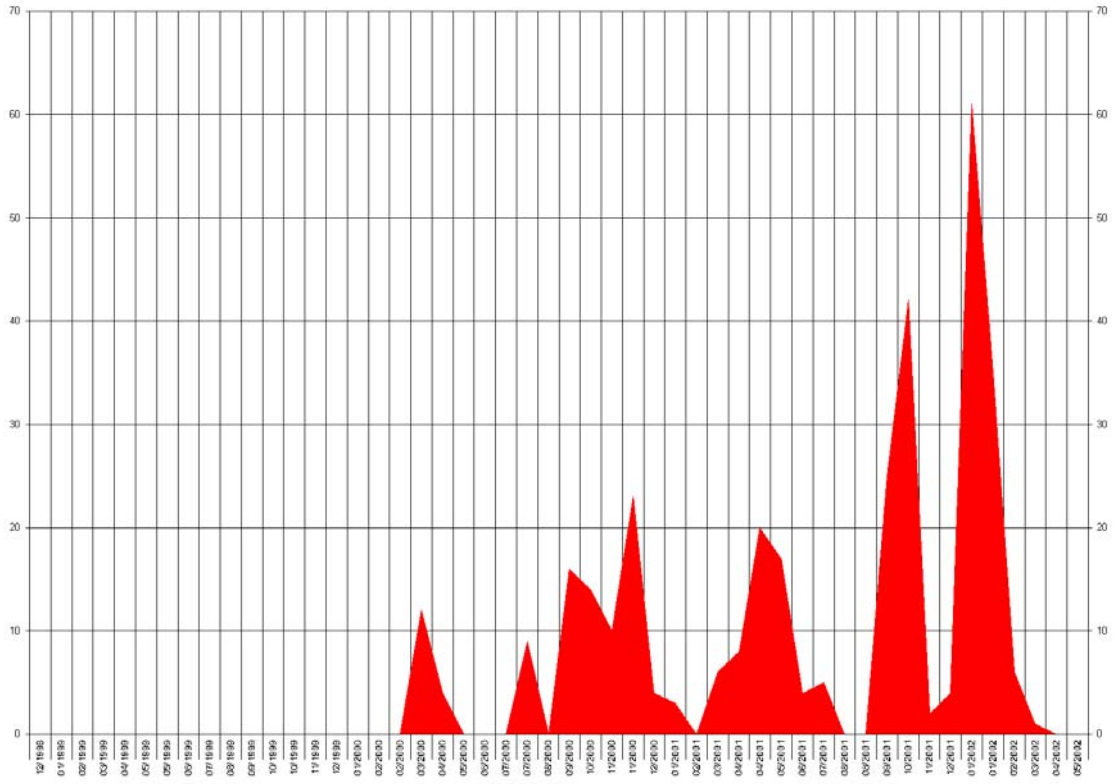

Figure 7 - Keyword tracing of "shred" (3D Area View)

Below is a step-by-step overview of performing a keyword trace.

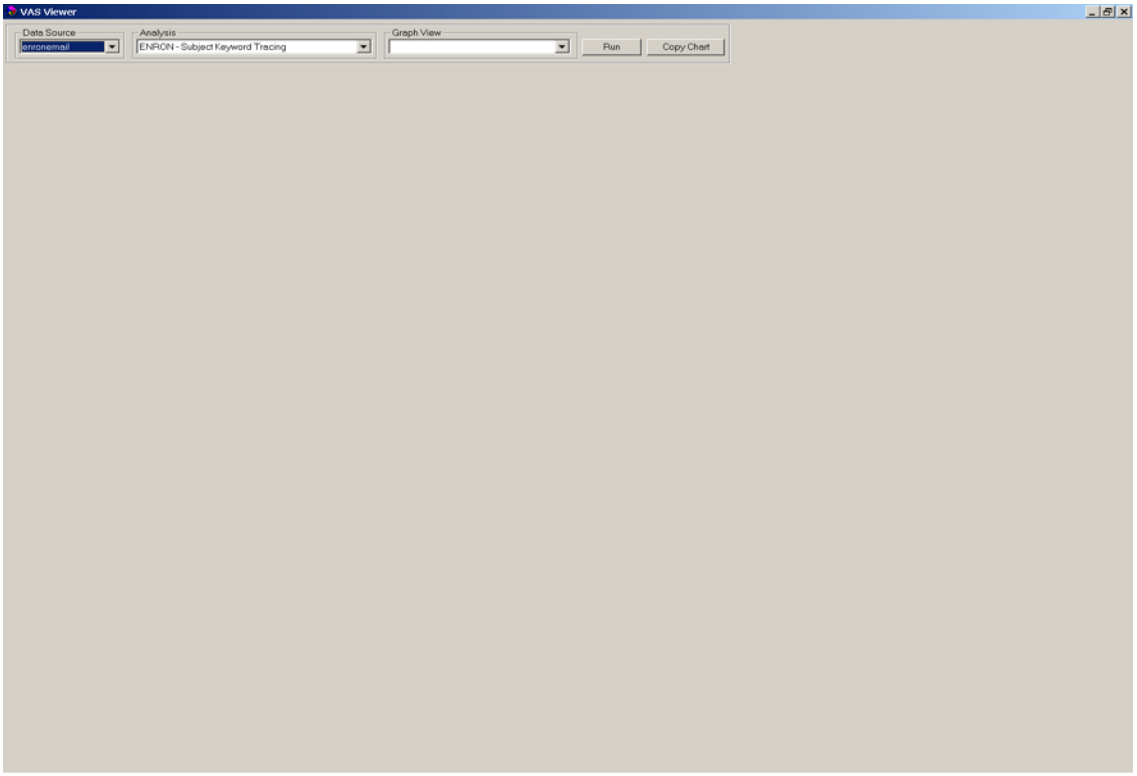

Figure 8 - Performing a Keyword Trace (Selecting Data Source)

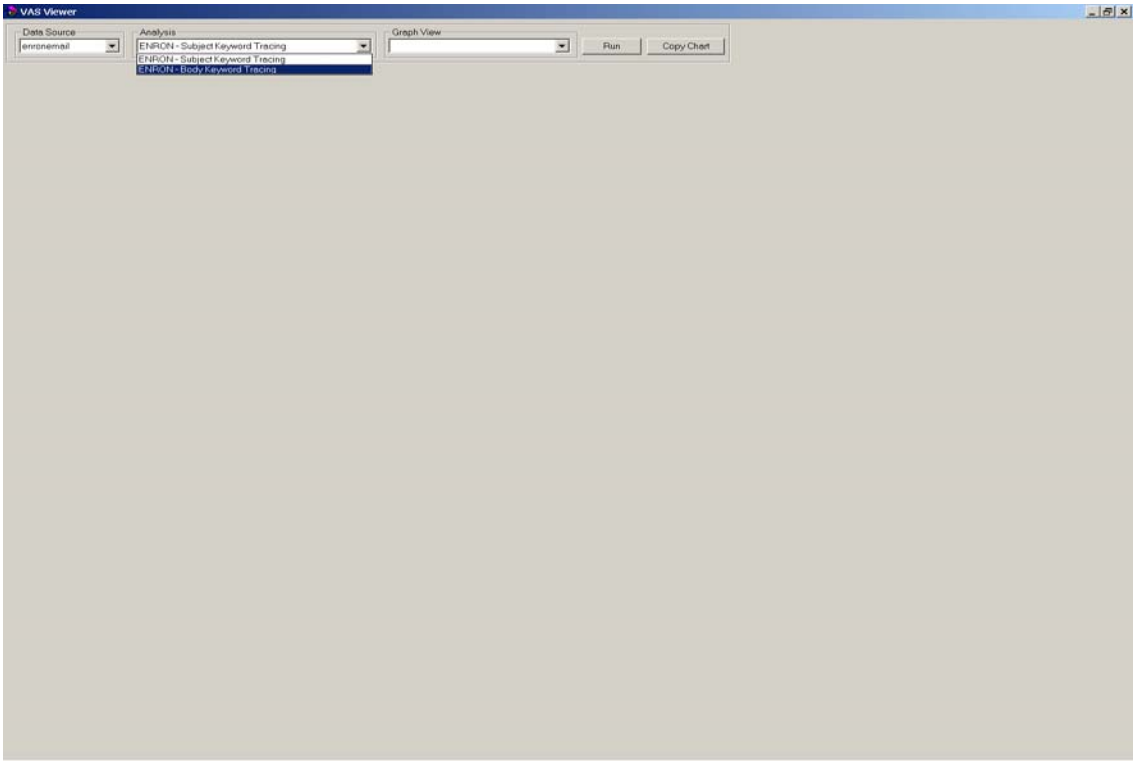

Figure 9 - - Performing a Keyword Trace (Selecting Analysis Type)

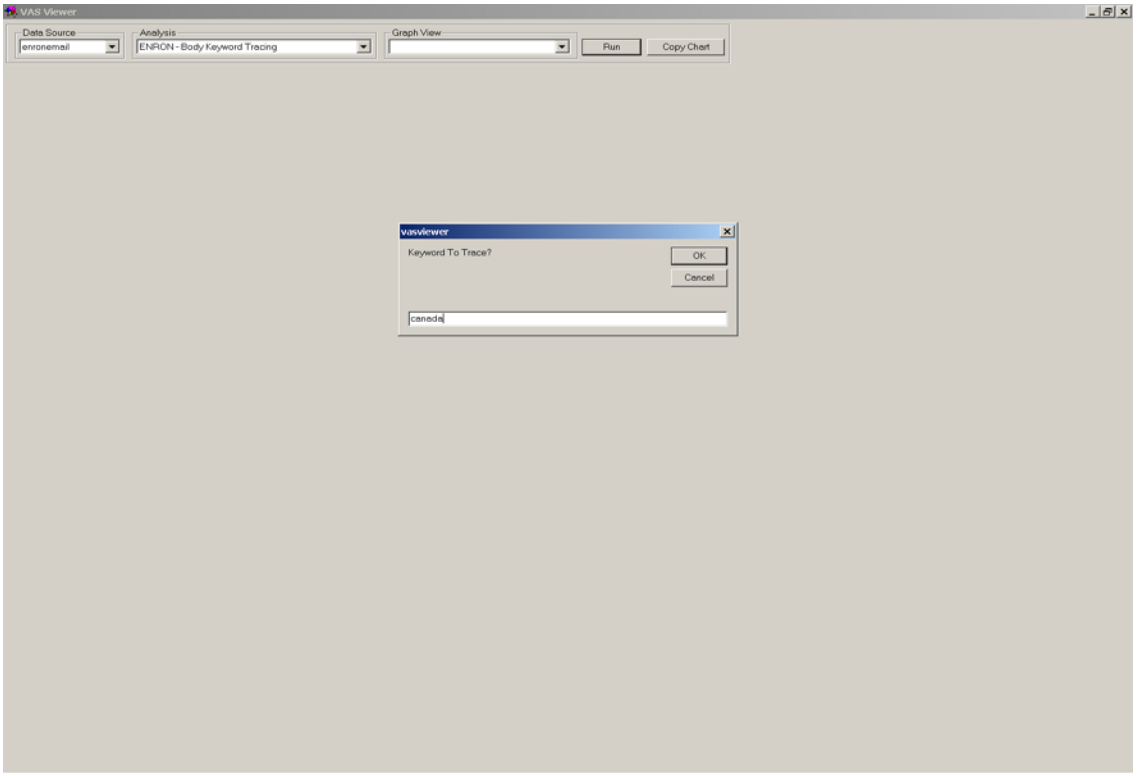

Figure 10 - Performing a Keyword Trace (Specifying Keyword)

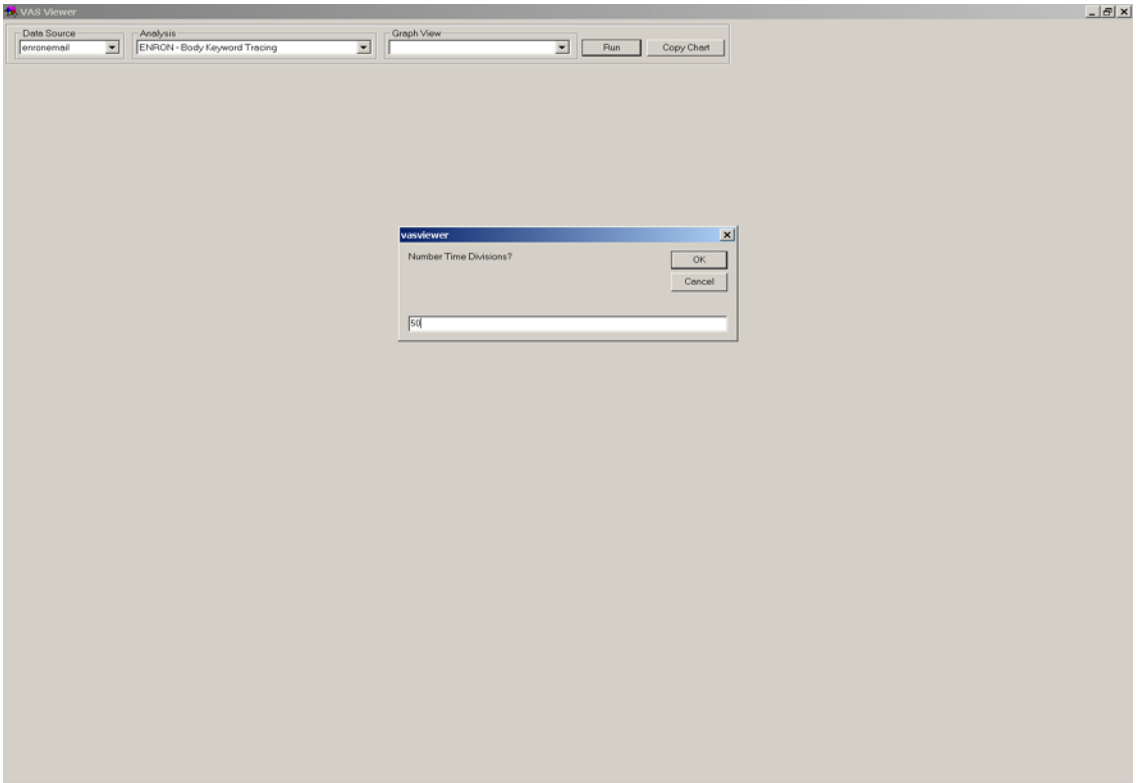

Figure 11 - Performing a Keyword Trace (Specifying Number of Timesteps)

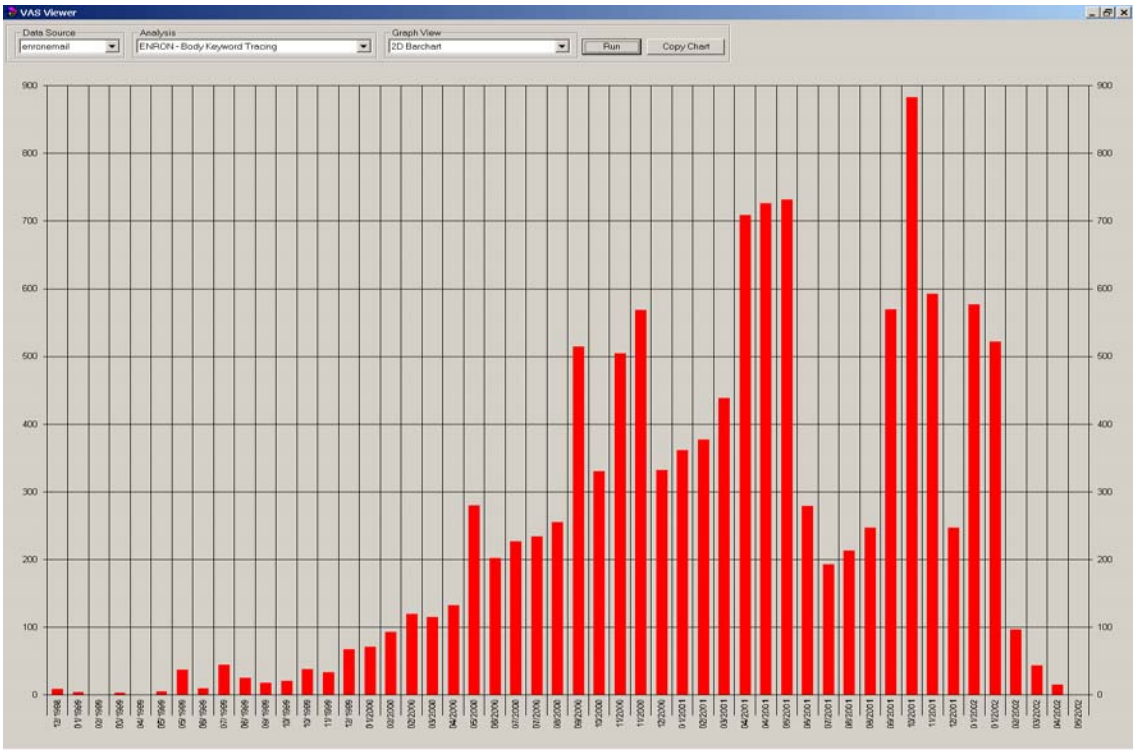

Figure 12 - Performing a Keyword Trace (Viewing Results)

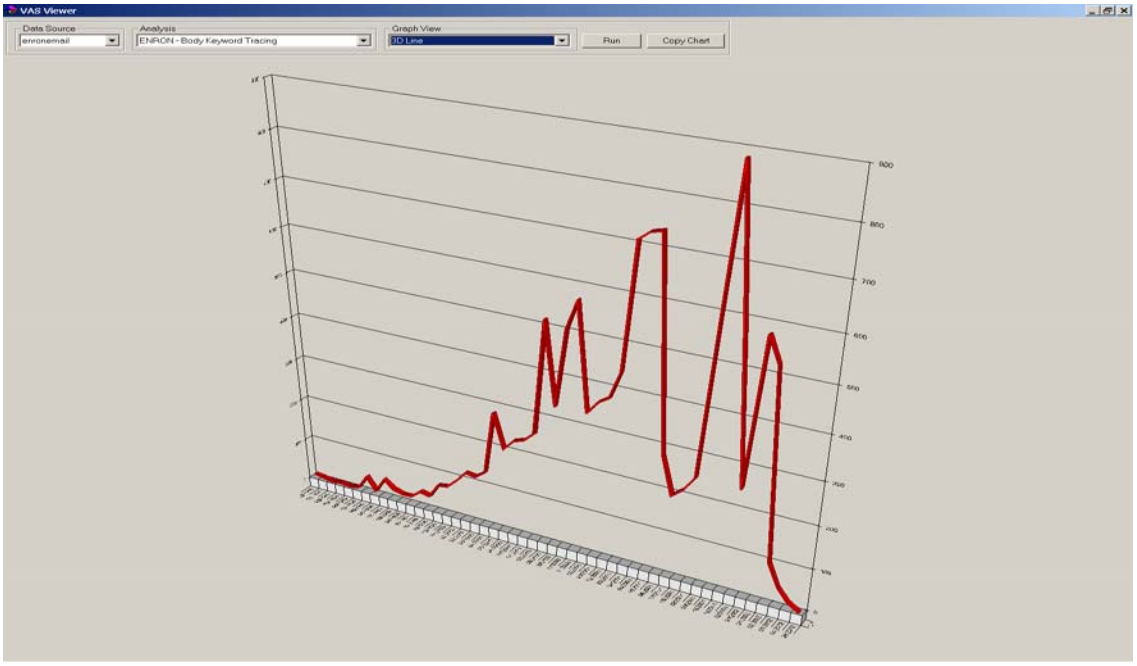

Figure 13 - Performing a Keyword Trace (Alternative View of Results)

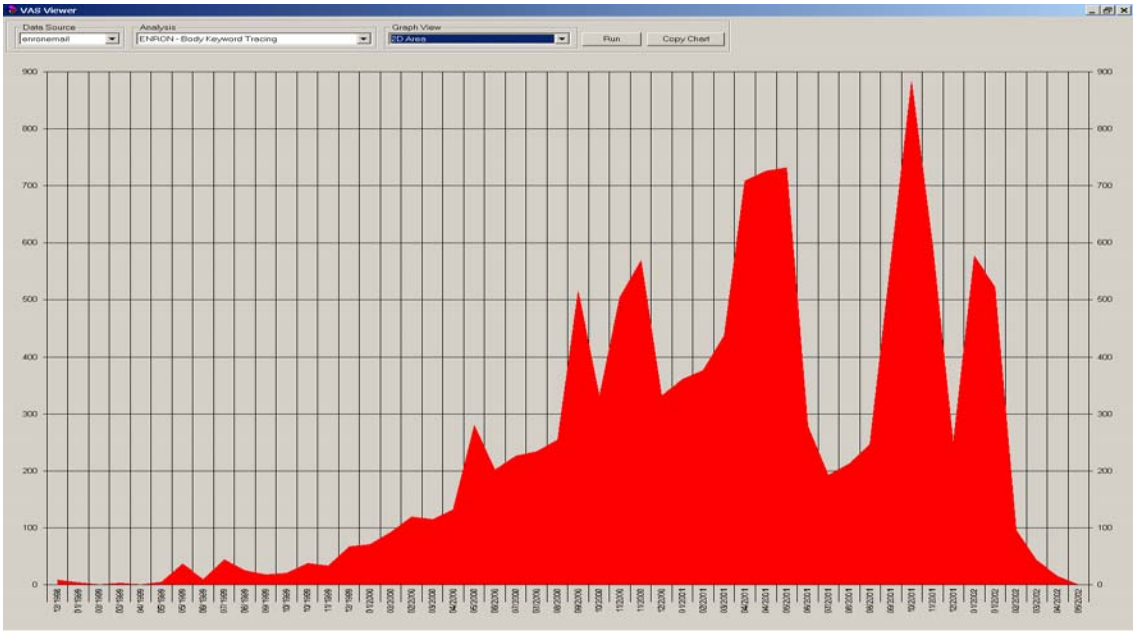

Figure 14 - Performing a Keyword Trace (Alternative View of Results)

In addition to keyword tracing, we also asked the system to identify the top thirty most prolific authors and the top thirty most common subject lines, seen in the graphs below.

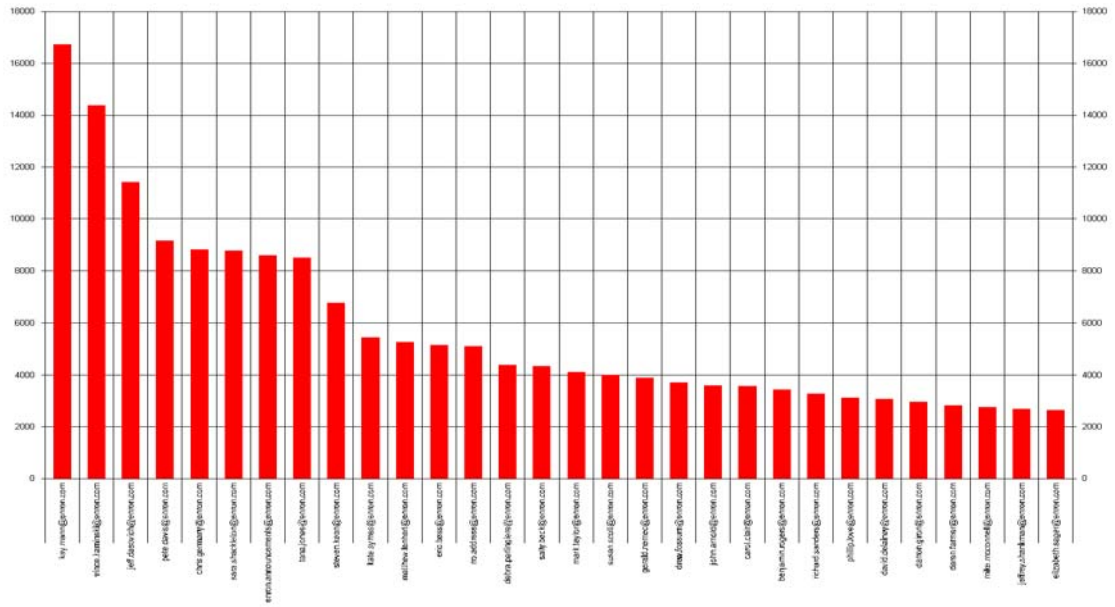

Figure 15 - Top 30 Most Prolific Authors

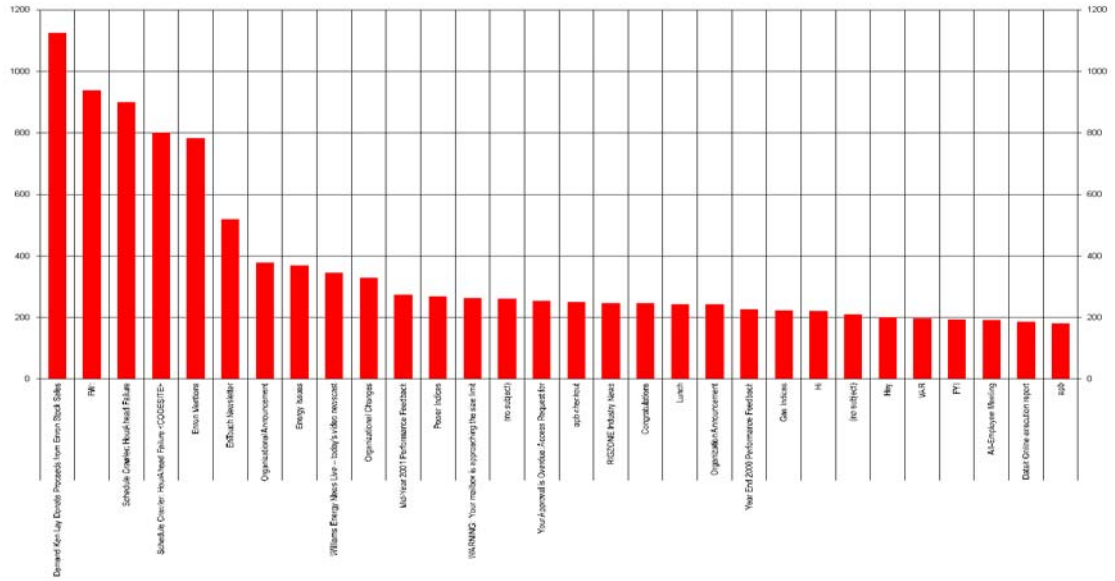

Figure 16 - Top 30 Most Common Subject Lines

NETWORK VISUALIZATIONS

The resulting structural network of the email collection was visualized on the NCSA Tiled Display Wall, a 40-projector rear-projected seamless display surface.

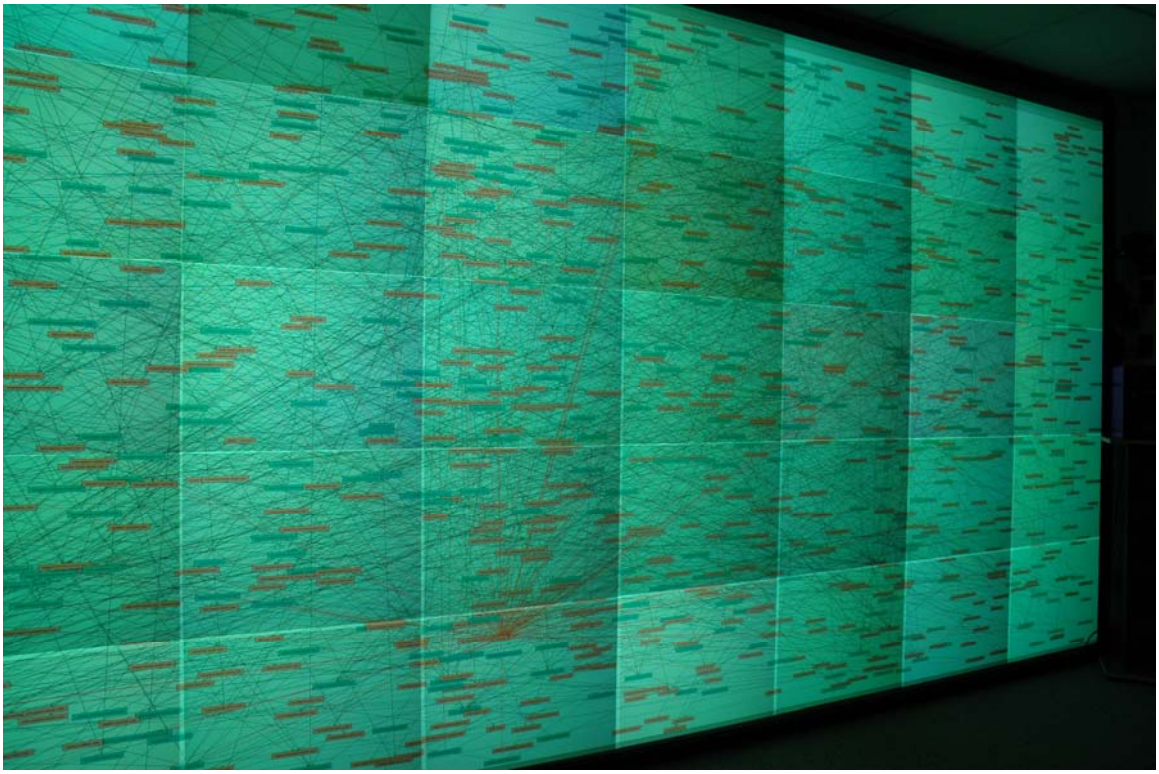

Figure 17 - Wide view of ENRON email network

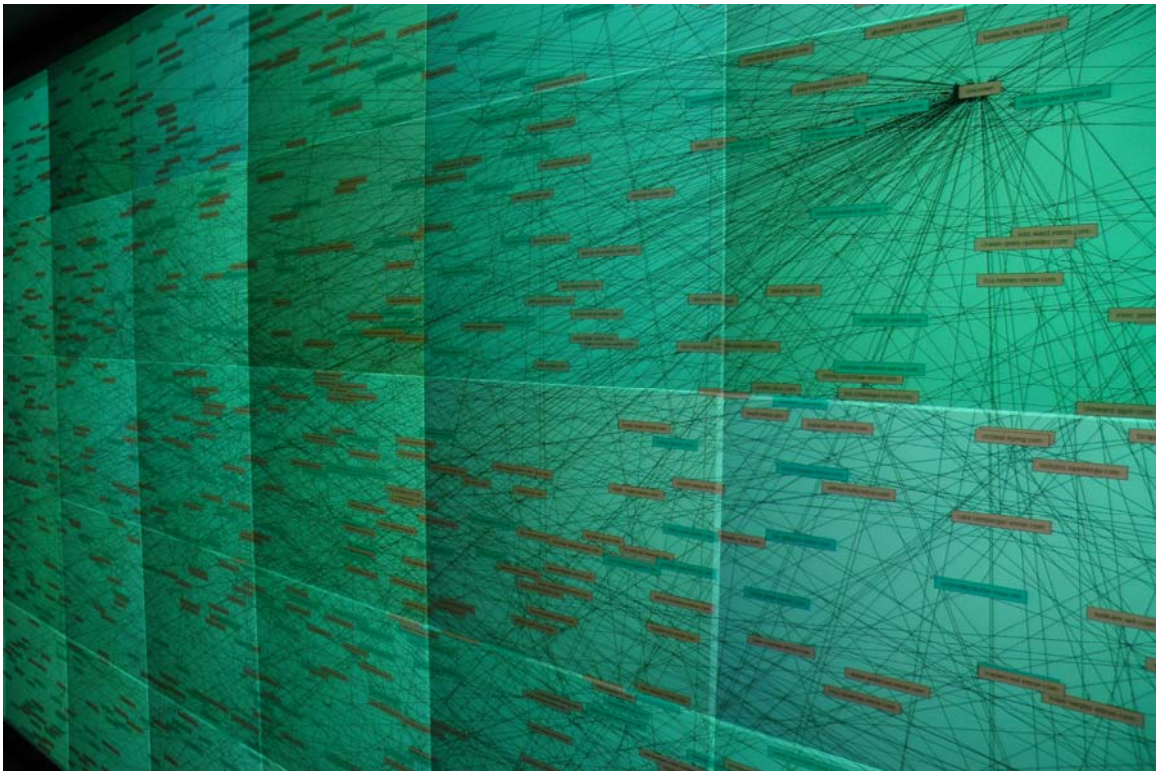

Figure 18 – Close-up of ENRON email network showing highly-connected node in upper-right

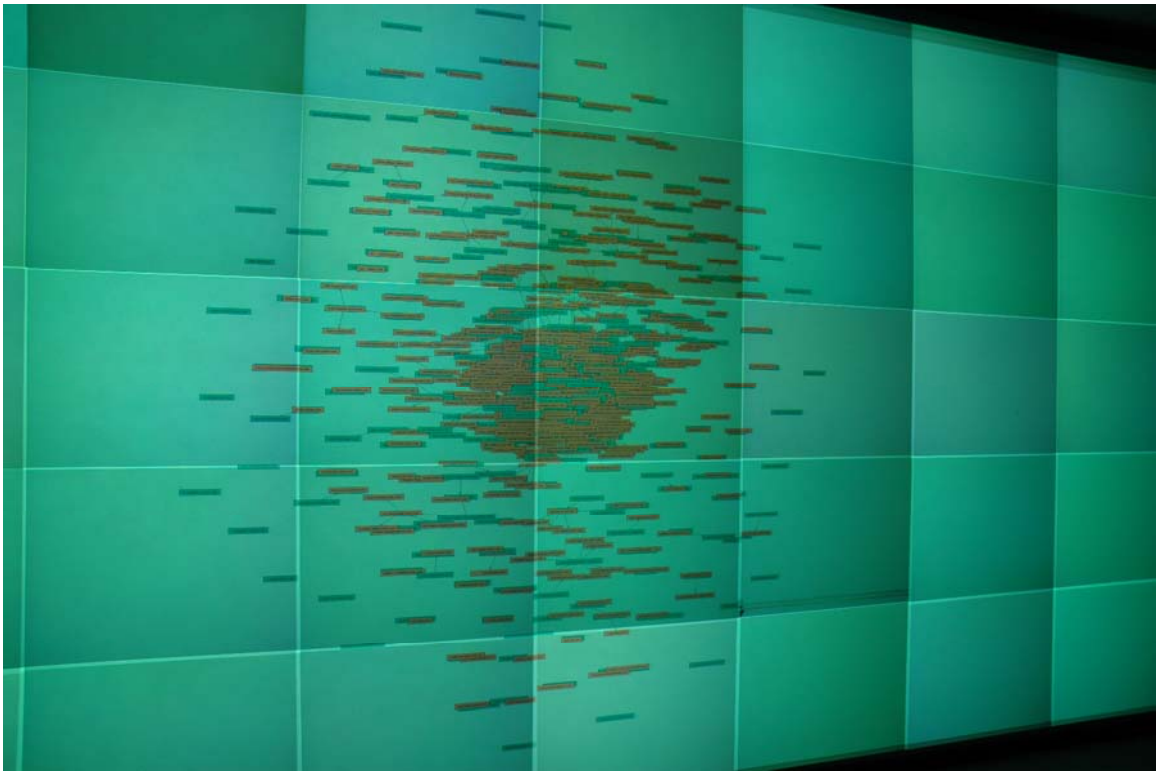

Figure 19 - Clustered network display of ENRON email network showing highly-connected interior nodes and loosely-connected external nodes