CS 224W Project Report: Topic Diffusion Under Crisis within the Enron Email Corpus

Evan Shieh, Ranajay Sen, Matt Anderson

I. Abstract

Following the 2001 Enron scandal, the Federal Energy Regulatory Commission (FERC) released a massive trove of emails - six years of correspondence to or from 151 Enron employees - that provided raw documentation of the leadup to and aftermath of what was the largest and most complex bankruptcy in American history. Though compelling for its singular nature alone (few such sets of email data are public), the Enron email dataset provides a special opportunity to study an organization under crisis. Building on prior work that narrowly examined variation in the structure of the Enron communication network over time, we have endeavored to enrich this analysis with metrics of topic and information diffusion, finding strong evidence for increased centralization/information siloing in organizations under crisis in addition to identifying topics that were talked about in isolated, private correspondences.

II. Introduction and Problem Description

In this paper, we are interested in studying how organizational information diffusion varies between regular and crisis conditions. The sanitized and deduplicated Enron email dataset published by Shetty and Adibi[1], consisting of 252,759 emails sent to or from 151 Enron employees between January 1998 and February 2004, presents a unique opportunity for this purpose: after reaching an all-time high stock price of $90 per share in August 2000, Enron saw crisis after crisis in 2001, including the sudden resignation of CEO Jeffrey Skilling in August, the launch of an SEC investigation in October, an unsuccessful buyout by rival firm Dynergy in November, and bankruptcy and mass layoffs in December. By sampling email communications from crisis and non-crisis periods of the whole dataset, we can study the effect of organizational crisis while controlling for a whole host of factors. Specifically, we can generate directed graphs of individuals (nodes) connected by emails (edges) and examine both network structure and (semantic) email content. Utilizing topic models that provide a rich natural language descriptor for email content, we can characterize edges by the topics of the emails that comprise them and nodes by the topics of their outgoing edges. Doing so allows us to rigorously test hypotheses about the behavior of organizations under crisis.

III. Related Work

In a paper titled “Communication Network Dynamics During Organizational Crisis”, Hossain et al.[2] set out to study a set of questions similar to ours; Hossain et al. sought to test five key propositions from the literature on sociology regarding the behavior of organizations under crisis: (1) that the most prominent actors will become central, (2) that the reciprocity of communication (the number of dyads in the graph) will increase, (3) that the transitivity of communication (the number of balanced triangles) will decrease, (4) that the

---

[1] Shetty, Jitesh, and Jafar Adibi. "The Enron email dataset database schema and brief statistical report." *Information Sciences Institute Technical Report, University of Southern California* 4 (2004).

[2] Liaquat Hossain, Shahriar Tanvir Murshed, Shahadat Uddin, Communication network dynamics during organizational crisis, Journal of Informetrics, Volume 7, Issue 1, January 2013, Pages 16-35, ISSN 1751-1577, http://dx.doi.org/10.1016/j.joi.2012.07.006. (http://www.sciencedirect.com/science/article/pii/S1751157712000570)

number of cliques will increase, and (5) that the organization will become more centralized.[3] Generating graphs for every one-month period in the Enron email dataset and measuring a number of network properties, Hossain et al. found patterns of variation that strongly supported hypotheses (1), (3), and (4) but much weaker signals concerning hypotheses (2) and (5). Although Hossain et al. performed a similar level of sanitization to that performed by Shetty and Adibi in the dataset we use, and although their regime of using a threshold on the number of emails between two accounts in a one-month period is similar to ours, it is notable that their analysis depended wholly on network structure, ignoring the semantic content of the emails and therefore the more qualitative aspects of the communication between individuals. By computing topic models for these emails (and therefore these nodes and edges), it is our hope to push beyond the limits of the work done by Hossain et al.

IV. Method

For the purposes of our paper, we settled on a methodology similar to Hossain et. al - from the whole six-year dataset, we extracted three one-month periods: October 2000 (period A), August 20001 (period B), and November 2001 (period C). The first period was our non-crisis control (Enron shares had hit their all-time high of $90/shar in August) and the latter two were both periods of turbulence (Jeffrey Skilling unexpectedly resigned as CEO on August 14th, two days before individuals voiced concerns about accounting practices internally, and November saw a steady slide in stock price as Enron failed to negotiate a buyout and filed for bankruptcy protection).

To generate the graphs for these one-month periods, we used a MySQL dump published by Shetty and Adibi[4] with all the information stored in four key tables: messages, recipients, email contacts, and references (to original quoted content). This solved a number of issues we had with other versions of the data set, since it had a thorough (though not full) entity resolution, allowed access to all the metadata of the data set, and was easily query-able. The ability to craft and use featureful SQL queries for the purposes of NLP and graph generation allowed us to quickly get back large sets of results. With our database setup and accessible to connection, our next step was to bring the data back to the graphical world. At this point, we created to develop a generator that allowed us to parameterize out the emails we wanted and create a family of graphs based on that parametrization. The generator allows you to specify [or not specify] a date range, a sender, a recipient, a keyword, a subject, or other relevant information. We then filter the nodes on that basis and create the graph of nodes of emails involved in that set. Each connection between two email nodes is a directed edge with a weight that represents the number of emails originating from that sender sent to that recipient. Our nodes also have metadata, embedding the associated email address and any other contact information we have. We also added the attributes of date and email subject/body to the edges for analysis in later steps.

At this point, we quickly realized how large and dense this email network dataset truly is, as without any further filtering most of the graphs we get back consist of tens of thousands of nodes and edges and millions of email bodies. That was far too much information to analyze meaningfully so we figured that we had to find a way to subsample the data such that fewer nodes, edges, and emails overall were returned and in doing so, we filtered out any noise. We decided on subsampling in three ways: 1. limiting the time range over which we considered all emails, 2. considering only edges between two individuals where at least 10 emails had been exchanged between the two and at least one email had been sent in either direction, and 3. randomly choosing a subset of all of the emails between two contacts as a representative set. After querying

[3] Ibid.
[4] Downloadable at https://www.cs.purdue.edu/homes/jpfeiff/enron.html

over the network with these thresholds, we were able to prune the appropriate nodes and edges from the graph and reduce the number of emails to create a much more tenable graph to use. The data was then output as a json file with a list of correspondences (between a sender and a receiver, each with a list of messages with relevant information), to be used in the LDA analysis.

A. Natural Language Analysis for Topic Mining

Despite being a highly unstructured form of data, emails represent a rich data source that can be used to characterize communication patterns. Compactly characterizing these emails can be used to infer deeper relationships between nodes in our graph, effectively attaching a rich set of attributes to each edge. For instance, the email that one sends to a business partner might be very different than the email that one sends to a secretary. Our approach to ascertain these differences utilizes Latent Dirichlet Allocation (LDA) as a robust, probabilistic topic model.

In broad strokes, topic models seek to discover "topics" that emerge from word tokens in a collection of documents. Latent Dirichlet Allocation is a generative, unsupervised algorithm that treats each item in a collection of documents as a mixture of latent (i.e. hidden) topics, each one drawing from a Dirichlet distribution. LDA is built upon a "bag-of-words" data model - the assumption that the order of words in each document can be neglected (and hence treats every document as a discrete histogram of word tokens).

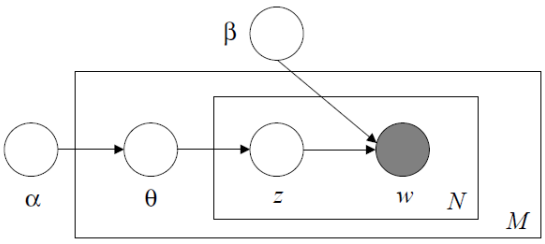


Fig 1. Probabilistic Graphical Model depicting Latent Dirichlet Allocation

LDA is a three-tier hierarchical Bayesian model, as outlined in Figure 1 above. It models the number of word tokens N in each document as generated by a Poisson distribution. For each document example, a topic prior probability distribution θ is learned by inference, where θ is a vector of dimension k (chosen by the user as the number of possible topics). θ follows a Dirichlet distribution parameterized by the model parameter α, as shown in below:

$$p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

Here, Γ is the Gamma function, which is used to normalize the product across exponentials of the elements of θ so that the probability distribution sums to 1. Given an inferred value of θ for each document, the model samples a topic z_n for each of the N words. Here, the topics $z \sim Multinomial(\theta)$ are treated as latent variables that are learned through inference on the marginal probability distribution below (across observed words w and model parameter β). We initialize $\alpha = k/50$ as a reasonable benchmark.

$$p(\mathbf{w} \mid \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta$$

Hence, LDA provides us with a model that can be learned to discover latent topics z based only on the co-occurrences and frequencies of word tokens (without even requiring their meaning). Performing parameter estimation on the conditional probability distribution above can be done using methods such as Gibbs sampling (which our implementation uses). Given LDA as the primary algorithm, our natural language analysis pipeline (implemented in MATLAB) is implemented in three broad stages:

1.) **Sanitize and pre-process emails.** For every email in our dataset, we preprocess the data in order to prepare it for robust topic analysis. Punctuation and extraneous numbers are removed, and every word is converted to lowercase (as words are treated as exchangeable in LDA). Common stop words in the English language are removed. We also Porter stem every word in order to perform consistent matching across writing styles.
2.) **Construct a Bag-of-Words (BOW) data model.** Word tokens in every document are as treated as entries in a histogram across dictionary values (where the dictionary is determined as the set of all unique stemmed words that appear in our email dataset).
3.) **Topic Analysis with Latent Dirichlet Allocation.** We assume an arbitrary number of topics to be learned, chosen to be $k = 100$ (which suited our purposes). A set of k topics is returned, along with a topic distribution for each document, which we normalize to sum to 1.
4.) **Semantic Similarity Metrics using Cosine Similarity**. Given a distribution across *k* topics, we then choose cosine similarity as a metric to compare how similar two distributions are. We prefer cosine similarity to other metrics like Euclidean distance because we are analyzing distributions, which all lie in the probability simplex.

As a proof of concept, we ran LDA (using Gibbs Sampling for inference) for 300 iterations in our milestone on 268 emails authored by Ken Lay across various time periods, and received the following sample topics (each shown below as a list of the words most likely to belong in that particular topic):

Topic 1: *'oil', 'gas', 'mena', 'report', 'petroleum', 'power', 'uk', 'algeria' …*
Topic 2: *'doc', 'enron', 'attached', 'joe', 'information', 'service', 'file', 'approved' …*
Topic 3: *'outlook', 'confirm', 'calendar', 'fleming', 'computer', 'rosalie', 'office' ...*

As seen by the examples above, each topic roughly pertains to a semantic category. We might even be able to infer, for instance, that Ken had a secretary named Rosalie (Rosalie Fleming was in fact Lay's assistant at Enron). Applied across several documents, this pipeline then gives us the ability to represent the semantic content of any given email as distribution of topics (approximated as a point in the *k*-simplex). Since each edge in our email network represents email exchange, we can average these topic distributions across every email pertaining to a given edge to pertain a topic distribution characterizing the interactions between any two nodes (i.e. people) in our network. This provides us with an incredibly rich attribute (depicted in Figure 2 below) that we can attach to each edge in our network, whose effect over time we can analyze using network analysis techniques.

Fig 2. Edge attributes learned using LDA. Each histogram is a distribution of word topics in the edges' emails.

B. Analysis Methods

1. Shortest Path Length

Now that we had a LDA vector associated with every edge, we now had a metric by which to compare graph edges, but still need a way to represent each node. At this point, we calculated the LDA vector of every node by averaging the LDA vectors of all of the outgoing edges of that node. In short, the semantic topic representation of any individual is based on all of the semantic topics in the emails that individual sends out. Now that we had the LDA of any particular node, we could find the semantic similarity between the topics any two nodes or individuals talked about using the cosine similarity metric detailed above. We then chose to compare the semantic similarity of two individuals with how separated they were in the graph. The metric we chose to represent the latter was the shortest path length between two nodes. We hypothesized that there should be a natural decrease in semantic similarity as the path distance from a node to another node increased, but the interesting information would be found in the differences of these distributions across time periods (in times of crisis and otherwise).

To do so, we generated the graphs for each time period A, B, and C, analyzed the LDA vector of each email, used that to generate the LDA vector for each edge. We then averaged the LDA vectors across outgoing edges to get the LDA vector of each node. For all pairs of nodes, we mapped the pairs to a store of the cosine semantic similarity of their respective LDA vectors. Then, we used snap.py functionality backed by a BFS to find the shortest path distances from any node to all other reachable nodes in the graph. Using this, we took each node, and made a distribution of the shortest path distance to other nodes and the average cosine similarity of all nodes that shortest path distance away. Now that we had a distribution of average cosine similarity vs shortest path distance, we averaged distributions across all nodes in that graph to generate the average distribution for that time period.

2. Bridge Edge

Our analysis on shortest path length led us to look into some metric to evaluate which the most important edges of the graphs, and focus on their LDA vectors. To do so, we were able to leverage the concept of bridge edges. Using the snap.py functionality to determine bridge edges, we were now aware of which pairs of individuals were responsible for communicating between different components in the graph. We wanted to see whether communication across these bridge edges were more or less different than across non-bridge-edges in times of crisis.

In order to determine this, once we generated graphs and LDA vectors for edges as above, we used snap.py functionality to find the bridge edges. We iterated over the bridge edges and for each node on a bridge edge, we compared how individual communicated across that bridge as opposed to all of his other contacts. To do so, we calculated the semantic similarity between the bridge edge and each other outgoing edges from that node on the bridge edge, and averaged these similarities to see how similar each bridge edge node was across the bridge vs not across the bridge edge. We then generated a distribution of these similarities for each time period, and compared the distributions across time periods to see if there were significant differences in times of crisis.

3. Community Comparison

Beyond semantic changes across shortest paths and bridge edges, we also strove to understand the variation in email semantics between and internal to communities in each of the three graphs. To do so, we computed communities in each of the graphs with Girvan-Newman community detection and found the average pairwise similarity between nodes in every pair of communities. By visualizing these averages in a

two-dimensional heatmap, we were able to directly compare community semantic self-similariy to inter-community semantic similarity.

V. Results and Findings

A. Analyzing Spread of Semantic Similarity (as a function of path length)

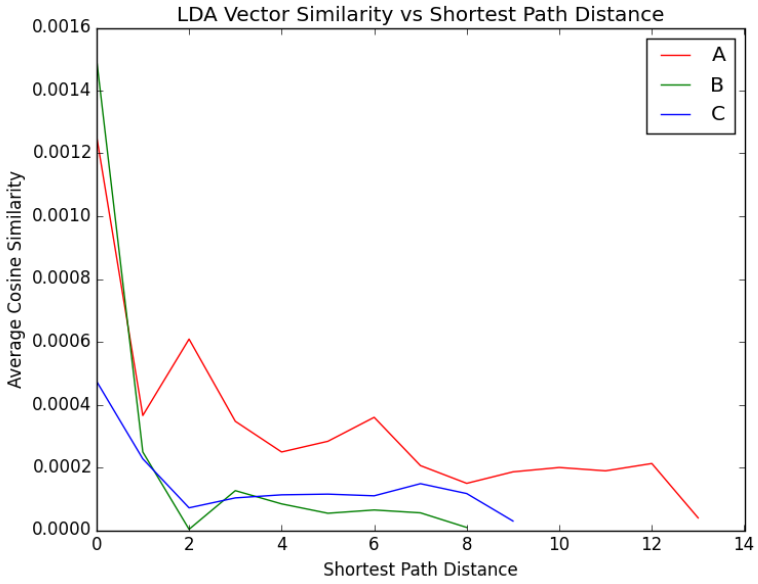


Fig 3. Semantic similarity generally decays as a function of node distance

The figure above depicts our analysis of how semantic topics diffuse across the graph, as a function of path length. The y-axis represents semantic similarity (more precisely computed as the cosine similarity between the LDA topic vectors for a pair of nodes), and the x-axis represents node distance (as measured by the length of the shortest path between two nodes). The three lines correspond to the diffusion patterns of each of the time periods analyzed (where A denotes the earliest period and C refers to the last period). A few observations stand out immediately. Firstly, the width of the graphs appear to be inversely proportional to the level of crisis. Graph A has a maximum path length of 13, whereas graph C has a maximum path length of 8. Furthermore, graph A has a more substantial long tail. That is, cohorts who are several hops away from the average node in graph A are more likely to have more similar communication patterns than cohorts who are several hops away from the average node in graph B (or C).

These two observations support the conclusion that email communications at Enron in earlier periods (October 2000) are more inclusive and more likely to support cross-pollination. In period A, co-workers who are 6 hops away in the email network are, on average, still talking about topics that are as similar as co-workers who are 1 hop away. By contrast, communication patterns at Enron during times of crisis become more siloed. In periods B and C, fewer long-range paths exist, and semantic similarity between nodes sharply decreases as path distance between the nodes increases.

When we conducted this analysis, we observed a strange artifact that was not explained by the trends mentioned above: similarity seems to behave strangely for nodes that are of path distance 2 apart (that is, the correspondent of a correspondent). In graph A, similarity seems to increase for node pairs of distance 2, whereas in graphs B and C, similarity seems to decrease for node pairs of distance 2 (as opposed to nodes who were distance 1 or 3 apart). After doing some deeper analysis, we discovered that nodes were more

likely to have neighbors who were 3 hops away than 2, for all graphs. This can be explained by a high proportion of bridge edges, as depicted below:

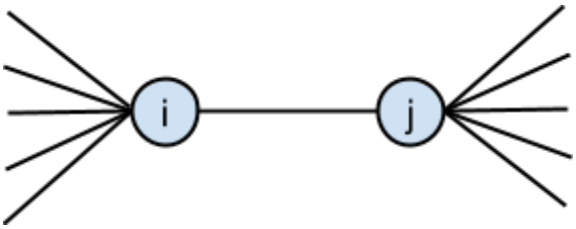


Fig 4. Depiction of bridge edges in our email communication network

Recall that by bridge edges, we mean edges whose removal would increase the number of separate connected components in the graph. If the articulation points *i* and *j* have high degree (say, $O(K)$), then we can observe that this structure produces more node pairs of distance 3 (roughly $O(K^2)$) than node pairs of distance 2 (roughly $O(K)$).

Bridges in an email communication network have an interesting set of possible corresponding interpretations in the corporate world. For instance, nodes *i* and *j* above could be managers whose individual subordinates only communicate across teams via their managers. Nodes *i* and *j* could represent acquaintances across separate departments who have developed friendships outside of work. Most compellingly, nodes *i* and *j* could refer to conspirators colluding from separate ends of the company. In any event, it is clear that bridge edges are a special subset of communications that are *isolated* in nature - the immediate peers of these bridge nodes do not communicate with each other. Hence, bridge edges represent an interesting way to measure how information and topics travel from one community to another, which we turn our analysis towards.

B. Analyzing Patterns of Stand-Alone Correspondences

When we analyze bridge edges, we are really analyzing patterns of communication that appear to be isolated within the network. Since the removal of a bridge separates one component into two smaller components, the flow of information that occurs between two persons along a bridge cannot be substituted by another relationship within the email network. Hence, communications along a bridge edge can be thought of as *stand-alone correspondences*.

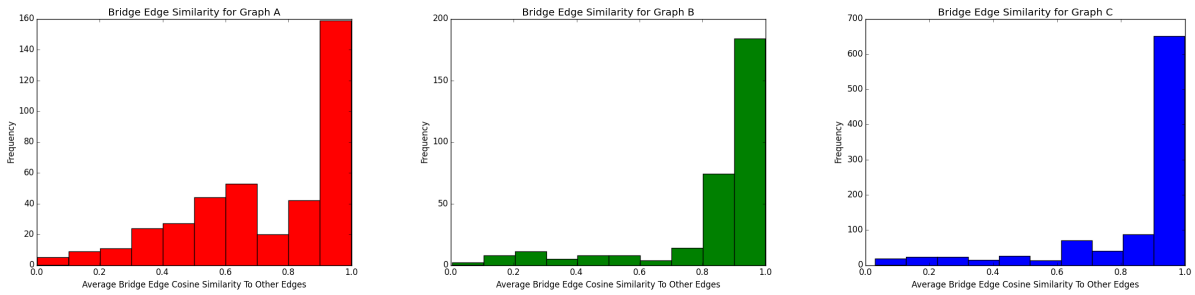


Fig 5. Similarity of Stand-Alone Correspondences and Regular Correspondences

The histograms above show the relative similarities between stand-alone correspondences and regular correspondences for bridge nodes in graphs A, B and C (depicted from left to right, respectively). The x-axis

represents the semantic similarity between the bridge edge and non-bridge edges for a particular bridge node, and the y-axis represents frequency. We can see that in all three graphs, the most modal similarity measures fall in the range 0.9 to 1.0, which means that the majority of stand-alone correspondences were extremely similar to regular correspondences. However, a good portion of stand-alone correspondences differed notably from regular correspondences, falling in the range 0-0.7 on a cosine similarity scale. We denote these types of correspondences to be *isolated correspondences*. At best, isolated correspondences represent casual conversation between two friends on the Enron network. At worst, isolated correspondences can be indicative of collusion between two conspirators, sharing different pieces of information than what they discuss with the rest of their immediate peers.

We note that the histogram changes in shape in a rather consistent pattern - isolated correspondences decrease sharply in frequency from graph A to C. Hence, during time periods of crisis, isolated correspondences were curtailed heavily, though not entirely.

We decided to analyze the most frequent topics of conversation in these isolated correspondences by looking at emails along bridge edges with a similarity score in the range 0.3 to 0.7, discluding emails that were either too commonplace (>= 0.8 similarity) or too off-topic (<= 0.2 - similarity). To determine these "most isolated topics", we determined the 3 topics most frequently found within isolated correspondences that did **not** appear in the 10 most frequent topics for regular correspondences. Our results are shown in the table below:

Period	Topic ID's	Frequencies (%)	Most Common Words
A	99, 92, 66	5.8, 4.9, 3.4	janette, time, legal, expenses, elbertson, cook, chaundra, department, moved
B	94, 25, 79	35.0, 13.0, 11.4	enrononline, report, summary, metafile, embedded, outside, ensure, decision
C	61, 80, 8	20.4, 14.4, 4.4	report, enrononline, transactions, outside, site, transaction, cera, nytimes

Fig 6. Most Common Topics Within Isolated Correspondences

The first column above denotes the time period of our analysis; the second column denotes the top 3 isolated correspondence topics. The third column denotes the relative frequencies of these topics - that is, the likelihood that any word in an isolated correspondence will be drawn from that particular topic. The fourth column is an aggregate list of the "top" words that are most likely to be drawn from the three listed topics in each row.

The topics in period A could perhaps be best categorized as "logistical", in dealing with time and expenses. However, a closer examination yields some insights. For instance, the word Janette was a top word for both topics 99 and 92, the two most frequent isolated topics for period A. Janette Elbertson was the secretary to Mark Haedicke, who is best known in this context as Enron's general legal counsel in North America. Haedicke was also alleged to have had a role in the Enron scandal - Salon reported in 2002 that Haedicke "is described … as sitting back and doing nothing when alarms were sounded about the controversial shell partnerships blamed for the company's implosion. Despite this nonchalance, Haedicke was compensated with $750,000 in bonuses last November. Haedicke is also about to leave Enron to work for UBS" [5].

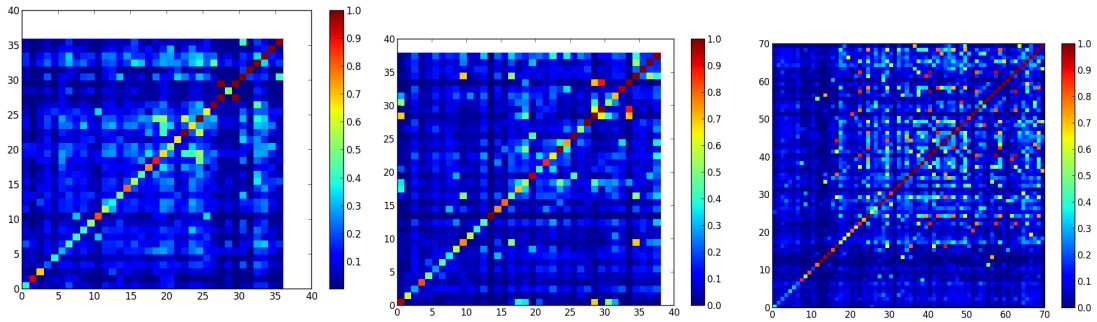
[5] Tapper, Jake. "Enron's last-minute bonus orgy". Salon, 2002. http://www.salon.com/2002/02/08/enron_bonuses/

While it's not definitively clear that this relationship between Haedicke and Janette's appearance in isolated communications is more than a coincidence, it is intriguing that Janette Elbertson, out of all secretaries at Enron, was most likely to be mentioned in (or be a part of) communications along isolated correspondences within the Enron email network during October of 2000.

Perhaps even more compelling is the presence of the token "enrononline" in both graphs B and C. For context, Enron Online was Enron's electronic commodity market (which opened in November of 1999) that was the first of its kind, allowing Enron to transact solely with participants in the global energy markets and eventually reaching up to $6 billion worth of transactions every day at its peak. Furthermore, New York Times reported in 2008 that Enron Online was a central mechanism by which Enron profited from what was known as the "Enron Loophole", the commodity futures act that "largely exempted the company from regulation of its energy trading on electronic commodity markets" [6].

In fact, in graph B, "enrononline" was a top word for all 3 most frequent topics to appear in isolated communications. Furthermore, these topics were discussed in isolated correspondences in period B with a far higher frequency than in any other time period (summing up to a 59.4% likelihood). To put that into context, these 3 topics numerically represent just 3% of all possible topics (100 in total). As a result, it's clear that Enron Online was a popular topic in stand-alone communications that was not frequently discussed in general communications during the months of August and November in 2001. At best, this represents gossip. At worst, this indicates collusion.

C. Community Comparison



From left to right, we have the heatmaps of average pairwise node similarities between communities numbered along the heatmap axes for graphs A, B, and C. Communities were sorted from largest to smallest, and communities beyond number 18 were seldom larger than 5 members. Of the regions of these heatmaps that represent meaningful communities, we notice that crisis does surprisingly little to increase inter-community similarity most colorful dots in heatmap C correspond to spurious similarities between single-member communities).

D. Conclusions

From the above analyses, it is clear that the pairing of semantic analysis with network analysis provides a compelling method for constructing how communication patterns evolve over time, especially in regards to crisis. Three salient months were chosen corresponding to periods of interest in Enron's history leading up to its collapse, capturing how email communications looked before the crisis as well as during the

[6] Lipton, Eric. "Gramm and the 'Enron Loophole'". New York Times, November 2008. http://www.nytimes.com/2008/11/17/business/17grammside.html

progression of the crisis. We've discovered that communications among teams became more siloed when Enron replaced Skilling as its CEO (period B) as well as when Enron was reaching bankruptcy (period C). Examining the notion of bridge edges allowed us to analyze *isolated correspondences*, which revealed which topics were most frequently discussed behind the backs of employee peers. Our analysis showed that Enron Online, a central cog in Enron's machined scandal, was a topic that was heavily discussed during these isolated correspondences that was not mentioned in regular conversations. Further work remains to be done to cement these analyses, but it these methods provide a compelling start point for using automated network analysis tools to examine patterns of communication.

VI. Team Contributions

A. Evan developed the email preprocessing and ability to do LDA analysis on any body of text, as well as the ability to analyze semantic topics across email subsets.
B. Matt did exploratory research on the Enron dataset and developed community comparison and generated the associated heat maps.
C. Ranajay created a graph generator to turn the dataset into a family of parameterizable graphs, and developed the shortest path and bridge edge analysis functionality.

VII. References

A. Liaquat Hossain, Shahriar Tanvir Murshed, Shahadat Uddin, Communication network dynamics during organizational crisis, Journal of Informetrics, Volume 7, Issue 1, January 2013, Pages 16-35, ISSN 1751-1577, http://dx.doi.org/10.1016/j.joi.2012.07.006. (http://www.sciencedirect.com/science/article/pii/S1751157712000570)
B. Lipton, Eric. "Gramm and the 'Enron Loophole'". New York Times, November 2008. http://www.nytimes.com/2008/11/17/business/17grammside.html
C. Shetty and Adibi schema: http://foreverdata.org/1009/Enron_Dataset_Report.pdf
D. Tapper, Jake. "Enron's last-minute bonus orgy". Salon, 2002. http://www.salon.com/2002/02/08/enron_bonuses/
E. Upgraded dump: https://www.cs.purdue.edu/homes/jpfeiff/enron.html