Crime Ring Analysis with Electric Networks

Team 17160

February 13, 2012

Abstract

In this report, we identify the participants in a corporate conspiracy plan using records of messages sent and received by employees of the company. We detail the motivation, mathematical details, robustness testing, strengths, and weaknesses of the two models we co-developed and used, and present and interpret their results. Both models use the senders and receivers of the messages, as well as their topic labels, but not timestamps or full-text content, which are not available to us.

Our first model is a vector space model which maps each of 83 nodes to a vector whose entries represent the number of communications between a node and each other node, as well as the number of communications involving a node and containing a specified topic (the vectors are of length 83 + 15 = 98). Within this model, we use the Euclidean norm as a distance measure, and show that k-means clustering into two clusters meaningfully separates known guilty and non-guilty employees when both node-node and node-topic interactions are considered, as well as when only node-topic interactions are considered, in both the case at hand as well as the example previous case. We also use cluster analysis to determine that, among all cases of repeated names, Beth is the most likely to be represented by two nodes. The clustering model does not assume known guilt or innocence a priori and doesn't strongly suggest any additional conspirators – however, its ability to separate known guilty and innocent employees granted credibility to the usage of linear spaces and equally weighted node-topic connections in our second, main model.

Our main model was an electrical circuit where each node and topic is represented by a vertex in the circuit graph, and people/topics are connected via a conductance proportional to the number of interactions made (these are the same as the corresponding entries in the vector space model). In this model, the people and topics given to be guilty were set to a reference voltage of 1 while those given to be innocent were grounded to 0. The voltages of unknown nodes could then be determined by solving the circuit in a DC setting. We chose this model because of its simplicity (to prevent overfitting), expressiveness, flexibility to accommodate new types of data and new problem times, and because the metrics it uses are supported by the clustering model.

We tested the robustness of our model by checking for leave-one-out discrepancies – that is, we experimentally ran the model once for each known conspirator, leaving that conspirator out of the set of vertexes held at 1V, and analyzed the consistency of these results with our main results. Our model held up to this analysis, showing that our results are not overly sensitive to the given known conspirators, and do not imply guilt by direct association to any individual. We likewise performed leave-one-out tests for each suspicious topic, and again the model held up. We ran the model as well as validation both for the case of Chris being innocent, as well as for the case of Christ being guilty (and topic 1 being suspicious).

We present our results and recommendations for who should be further investigated.

Contents

1 Problem Statement

In this problem, we aim to uncover the participants in a corporate conspiracy plan using records of messages sent and received by employees of the company. We are not given access to the full-text content of the messages – however, the messages are pre-labeled using 15 general conversation topics, the descriptions of which are given. Using this information, as well as a given set of known conspirators and known non-conspirators, we aim to identify a set of likely conspirators among the remaining employees. Of particular interest are three of the company's senior members: Gretchen, Jerome, and Delores.

2 Approach Philosophy

The development of our model was guided by a few basic philosophies, which arise from the legal nature of the problem and the nature and scope of the example given.

First, we wanted to avoid overfitting our data. Specifically, we determined that the general process of:

- Coming up with a generic model
- Fitting its parameters to best explain the given example
- Running the model with those parameters on the case at hand

would not be an effective approach. For example, a model with three weight parameters, tested at 11 values each (say 0 to 1 in increments of 0.1) would result in a search space of 1331 models. Among 10 people in the training example, there are 1024 possible combinations of guilty and innocent people. With no other assumptions, we would expect $\frac{1331}{1024} > 1$ set of three parameters to happen to match the results of the given example, regardless of whether or not those parameters result in a good general model.

Instead, we chose to develop two simple concurrent models that support each other. Our main model is an electric network model that assumes the guilt of those given to be guilty and flags topics given to be suspicious. To support this model and give credibility to the metrics, choices of fixed parameters, and data assumptions it uses, we developed an auxiliary clustering model. Both of these models are discussed in detail in sections 4 and 5.

Second, we wanted to avoid models that implied any level of presumption of guilt in any intermediate term (e.g. conditional probability terms in an intermediate iteration of a PageRank-like algorithm). Because the problem statement is ambiguous about the legal proceedings themselves, and of exactly why the data is abridged the way it is, we wanted to avoid the chance of using any intermediate assumptions (even in an iterative calculation) that could could later undermine the admissibility of subsequent evidence collected on the basis of our recommendations (for example, a violation of probable cause).

Third, we understand that any algorithmic approach on the basis of communication data risks implies some degree of guilt by association. In order to protect the credibility of our work, we used leave-one-out testing as a measure of the robustness of our results – that is, for each person known to be involved in the conspiracy, we ran the model without assuming them guilty to ensure any of our results are not too dependent on associations with any particular individual being assumed guilty.

3 Data Interpretation and Assumptions

Our dataset consisted of 400 message headers (i.e. sender-receiver pairs) sent between 83 nodes, representing employees within the company. 7 (or 8, if counting Chris) employees were known to be tied to the conspiracy. The messages were additionally pre-categorized into 15 topics, 3 of which were known to be related to the conspiracy plan. Some messages contained multiple topics – 36 messages contained (exactly) 2 categories, and 11 contained 3 categories. Most of the 15 topics were associated with between 20 and 40 unique messages. We were not given the actual message contents.

318 out of the 400 messages were sent between a unique pair of employees, which unfortunately makes it difficult to compare the relative frequencies of contact between different pairs of individuals, as there were very few instances of a person sending more than one message to another individual. This made it difficult to meaningfully consider the communication frequencies between pairs of nodes without overfitting the data (we worked around this by considering node-to-topic connections in our clustering and electric network models).

There were two main ambiguities within our data. First, there were 5 employee names that corresponded to 2 nodes each. These names include Gretchen and Jerome, the names of two senior managers in the company. In each of these cases, it was unclear whether or not the two nodes were different communication nodes (e.g. different cell phones) that belonged to the same employee, or whether the two nodes represented different employees with the same name. We used cluster analysis to predict which such pairs of nodes to regard as single individuals.

The second main ambiguity was a single message (line 215 in Topics.xls) that contained three topics, one of which was marked 18, an undefined topic number, and a clear error in the data. We did not find any conclusive clustering evidence in favor of classifying this topic among any of the 13 possible third-topics, and threw the value out (e.g. we only associated the message with its two valid topics).

4 Clustering Model

In this section we set up a model that splits workers into groups based on interaction with others and with topics. Rather than use this as our main model for ranking the employees in terms of guilt, we use it to answer some preliminary questions such as whether or not duplicate names refer to the same person and how to exactly use the provided message data in our main model.

4.1 Model Description

We use k-means clustering to split the the employees into k disjoint sets based on the people and/or topics that an individual interacted with. In k-means clustering, the elements of the set are typically viewed as vectors, so the clustering essentially partitions the set of vectors into k subsets, where any vector $\mathbf{v}$ in subset S_i is closer to the centroid of S_i than it is to any $S_j \neq S_i$. We will take euclidean distance as the measure of closeness (which is quite standard). In our model, the rows of the 83×98 conductance matrix C, where C_{ij} equals the number of times person i interacted with person or topic j, are taken to be the vectors. (Note that we set C_{ii} to be the average of all of the other entries of column i.) For more details on C, please refer to the section on the circuit model. The intuition behind this approach is that if two people talk to the same people about the same topics, they should belong in the same cluster. The algorithm used proceeds as follows:

1. k vectors are randomly chosen from the 83 rows of C. These k vectors anchor k subsets of the row vectors of C.
2. Each row vector of C is assigned to the subset anchored by the vector closest to it.
3. Once all vectors are assigned, the centroid (average) of each subset of vectors is taken to be the new anchor.
4. Steps (2) and (3) are repeated until the assignments of vectors no longer change.

Note that this algorithm provides a fast local solution to an otherwise NP-hard problem of finding the global optimum (with respect to minimizing the sum of the length of the difference between each vector and the anchor of the subset to which it belongs). Therefore, the resulting clusters for a fixed k and C are not always the same; however, in all of the results below, we run the algorithm enough times to determine the most probable clustering.

4.2 Figuring Out Duplicate Names

Since there are 83 names with 5 repeats in the list of names and 82 employees at the company (given in the problem statement), exactly one of these duplicate names must refer to the same person. One use of clustering is to determine which name this is. Using our model, the two names that refer to the same person should be likely to lie in the same cluster since a unique person will usually interact with the same group of people about the same topics. On the other hand, two distinct workers who share a name have no reason to share similar interactions. Thus we look at whether or not each pair of repeated names lies in the same cluster for $k = 2$ through $k = 5$. For example, does node 16 Jerome fall into the same cluster as node 34 Jerome, etc. The results are shown in Table 1.

We note that Elsie and Jerome are in different clusters 3 out of 4 times, so these two names are likely to represent four distinct people. To figure out which of Gretchen, Beth, or Neal is redundant, we can look at the number of interactions between the two Gretchens, the two Beths and the two Neals. Node 4 Gretchen communicated with node 32 Gretchen once, node 17 Neal communicated with node 31 Neal once, and node 14 Beth did not communicate

k	**2**	**3**	**4**	**5**
Elsie	same	different	different	different
Jerome	different	different	different	same
Gretchen	same	same	different	same
Beth	same	same	same	same
Neal	same	same	same	different

Table 1: The table below shows whether or not each pair of duplicate names are grouped within the same cluster for four values of k.

with node 38 Beth. Given that it is unlikely that a person would message him/herself, we conclude that the two Beths refer to the same person and will merge their message history for the remainder of our solution.

Now that the two Elsies have been deemed different people, the question arises of which one is known to be guilty (since we are given that Elsie is a conspirator). This is another question we can answer with our clustering model, because for $k = 3, 4$, and 5, node 7 Elsie lies in a cluster of four with three other known conspirators. Meanwhile, node 37 Elsie lies in a cluster of 38 with 3 known conspirators, a cluster of 10 with 1 conspirator, and a cluster of 30 with 2 conspirators for these same k values. Thus, for the remainder of the problem, we will take node 7 Elsie as the given conspirator and treat node 27 Elsie as an unknown.

4.3 Validating Parameters

Finally, we use our clustering model to validate the parameters chosen for the main circuit model described in the next section. In the circuit model, we have the flexibility of tuning the relative weights of people-to-people interactions and people-to-topic interactions. While people-to-people interactions are important for determining whether or not two Jeromes are the same person, they become less relevant when it comes to determining the probability of a person being guilty. One reason for this is that a person may interact frequently with their coworker or boss who happens to be a conspirator; however, if these interactions are about sports or job-related matters, this person cannot be implicated. On the other hand, talking about a conspiratorial topic with anyone can be enough to warrant investigation. Another reason to place less emphasis on people-to-people interactions is that connections between people and topics inherently reflect (to some degree) connections between people. For example, if Anne and Bob talk to each other about topic X, their connection to each other is reflected through their connections to X.

Given these logical reasons for using people-to-topic interactions as the basis of our main circuit model, we would like to back them up with some objective results, which is where the clustering model comes in. We can examine the results of clustering people according to their interactions with other people (columns 1 through 83 in C), their interactions with topics alone (columns 84 through 98 in C), or a combination of both (columns 1 through 98 in C). A reasonable measure of effectiveness is looking at whether or not the people we

	Cluster 1	Cluster 2
p2p	5 guilty, 2 innocent [7 total]	2 guilty, 6 innocent [76 total]
p2t	7 guilty, 3 innocent [42 total]	0 guilty, 5 innocent [41 total]
both	6 guilty, 3 innocent [29 total]	1 guilty, 5 innocent [54 total]

Table 2: The composition of clusters for $k = 2$ are shown with respect to the number of known guilty and innocent people in each cluster. We examine clustering by people to people (p2p), people to topics (p2t), and both.

know to be guilty get clustered together. For $k = 2$ clusters, we get the results shown in Table 2.

Thus, we can see that using people-to-topic connections alone produces the most desirable clusters in terms of grouping guilty people together, which gives us more reason to use this paradigm as the basis for the circuit model.

5 Electric Network Model

5.1 Introduction and Motivation

The electric network model constructs a graph whose vertexes consist of nodes (in the company social graph) and topics of conversation, and assigns a voltage potential to each vertex representing how suspicious it is of conspiracy-related activity. Suspicious topics and known conspirators are held at a potential of 1V, and known non-conspirators are held at a potential of OV. More vertexes (particularly those representing nodes) are taken to be more suspicious of guilty activity if they have a higher electric potential.

Edges in the graph contain resistors, which represent the distance between the connected vertexes. We analyze these edges using conductance (inverse of resistance), which is 0 for unconnected vertexes, and increases based on co-occurring activity between the vertexes.

As per Random Walks and Electric Networks (Doyle and Snell), every Markovian random walk is equivalent to a closely related electric network of the same topology. Thus, using electric networks gives us all the expressive power that PageRank or related algorithms can provide. However, electric networks do not require the intuitively unwarranted normalization assumption, increasing their expressive power (see Strengths and Weaknesses of the Model, later in this section). Moreover, due to not dealing explicitly with probabilities conditional on guilt in any intermediate calculations, we avoid violating our approach philosophy. Finally, the ability to fix the potential of nodes and topics makes it easy to incorporate the given information on known participants and onparticipants in the conspiracy, and known suspicious topics.

The electric network model satisfies the triangle inequality – that is, the equivalent resistance between any two nodes does not exceed the sum of the resistances along any path between the nodes. This makes it a meaningful model to analyze interactions. Finally, the model is highly applicable (see Applicability to Generic Agent-Ranking Problems later in this section) and extendable (see Section 7).

This section lays out the details of the electric network model, its structure, setup, assumptions, parameters, strengths, and weaknesses.

5.2 Bipartite Graph Circuit Setup

5.2.1 Structure and Definitions

Our bipartite graph circuit G = (N, T, E) contains vertexes of two classes – nodes (N) of the employee network and topics (T) of messages. Each edge (in E) connects a node $n \in N$ to a topic $t \in T$. The weight of an edge is its electric conductance (the inverse of resistance) – a node and topic are closer in electric potential if the node is involved in more communications that contain that topic.

In the sections below, we will additionally use the following notation:

- M refers to the set of messages as M
- $\text{tx}(m)$ denotes the node which transmitted message m
- $\text{rx}(m)$ denotes the node which received message m
- $\text{topics}(m) \subset T$ refers to the set of topics contained in a message m
- $\text{refs}(t) \subset N$ refers to the set of nodes referenced within the summary of topic t.
- $\mathbf{I}_S$ is the indicator function of a set S.

5.2.2 Node to Topic Conductivity

The connectivity between a node vertex and a topic vertex is a measure of how closely the node is associated with the topic. We measured this by counting the number of times a person sent or received a message using a particular topic. These values coincide exactly with the node to topic values in the vector space model used for clustering in Section 4.

We also explored considering whether or not the topic summary references the particular person (weighted by the number of messages containing this topic). The general form of the conductivity between node n_i and topic t_j is given as:

$$c_{n_i,t_j} = \sum_{m \in M} \mathbf{I}_{\text{topics}(m)}(t_j) \times \left(\begin{cases} \alpha & \text{if } \text{tx}(m) = n_i \\ \beta & \text{if } \text{rx}(m) = n_i \\ 0 & \text{otherwise} \end{cases} + \begin{cases} \gamma & \text{if } n_i \in \text{refs}(t_j) \\ 0 & \text{otherwise} \end{cases} \right)$$

In practice, we found no compelling reason to treat sent and received messages differently – due to the success of equal weighting in clustering, we simply used $(\alpha, \beta) = (1, 1)$.

We considered using a nonzero γ to take into account messages containing topics which themselves referred to the node in question, but of which the node was neither a sender nor receiver. We decided to keep $\gamma = 0$ due to the risk of overfitting the data (especially since a few node to topic connections would be affected immensely, e.g. Paige to one of several topics mentioning her). We also acknowledged that messages with positive or negative sentiments toward a node would require different treatment. In Section 7, we discuss the possibility of using text sentiment analysis of individual messages to determine whether or not a node views another node favorably, and using that as a factor in node to node distance.

The particular node to topic conductances we used to obtain the data we present is given by:

$$\boxed{c_{n_i,t_j} = \sum_{m \in M} \mathbf{I}_{\text{topics}(m)}(t_j) \times \left(\begin{cases} 1 & \text{if } n_i \in \{\text{tx}(m), \text{rx}(m)\} \\ 0 & \text{otherwise} \end{cases} \right)}$$

which is the number of messages either transmitted or received by n_i in which topic t_j was discussed.

5.2.3 Node to Node Conductivity

As the bipartite graph structure suggests, we did not utilize node to node or topic to topic connections. However, they are natural possible additions to our model.

The connectivity between two nodes is a measure of the strength of *direct* connections between those two nodes. We measured this by counting the number of messages passed between those two nodes:

$$c_{n_i,n_j} = \sum_{m \in M} \begin{cases} \delta & \text{if } \text{tx}(m) = n_i \ \wedge \ \text{rx}(m) = n_j \\ \varepsilon & \text{if } \text{tx}(m) = n_j \ \wedge \ \text{rx}(m) = n_i \\ 0 & \text{otherwise} \end{cases}$$

As with node to topic conductivity, we did not distinguish between sent and received messages. We used $(\delta, \varepsilon) = (1, 1)$ for clustering to explore the repeated names, but used $(0, 0)$ for clustering based classification due to its superior performance and its simplicity (it removes a parameter from the model). We effectively used $(0, 0)$ in the bipartite graph circuit here – no direct connectivity between nodes implies a conductivity of 0.

One important thing to note is that, even in the bipartite model, a new message between n_i and n_j will increase the conductivities between those nodes and the topic nodes corresponding to the message. This in turn increases the equivalent conductivity between the two nodes' vertexes.

Also, as mentioned in Section 7, there is scope for introducing direct node-to-node conductivities to represent positive of negative sentiments between the nodes on the basis of full-text message contents.

5.2.4 Topic to Topic Conductivity

We considered adding direct conductivity between two topic vertexes based on their co-occurrence in messages. We decided against this for two reasons. First, there were too few messages with multiple topics for this to be meaningful. Second, we wanted to avoid any material difference between splitting a message into multiple single-topic messages.

Note that, if two topics are frequently discussed by the same nodes, the equivalent conductivity between the two topic vertexes will anyway increase.

In Section 7, we discuss the possibility of adding conductivity between topics that are materially similar using network semantic analysis, latent semantic analysis, and text analysis.

5.3 Applicability to Generic Agent-Ranking Problems

Our circuit model, as designed, is very flexible. It accommodates analysis of any problem in which agents are to be ranked by their network proximity to one set of known “positive” agents than a given set of “negative” agents using their proximity to each property in a set of properties of interest.

In this case, the set of positive agents is the set of known conspirator nodes, the negative agents are the known non-conspirator nodes, and the properties we examine are interest in the fifteen denoted topics.

We could just as well use this network to identify infected cells in a biological network. One could imagine the nodes representing cells (or, for tractability, regions of cells), and the properties being various chemical and biological properties of interest, such as the main infection, various other infections, particular cell traits, etc.

Probabilities of each cell’s predisposability to the given properties may represent the pairwise conductances between cell vertexes and property vertexes – a higher probability represents a higher disposition, which consistently results in increased conductivity.

We may also include node-to-node connectivities to represent the physical proximity of the nodes. Factors that determine these conductances may include effective distance (for example, a conductance being an exponential decay in effective distance), where effective distance takes into account the physical distance, as well as barriers between the cells, proximity to common blood vessels, etc. This information can be gleaned from images of the system.

The biological model suggested here is certainly more complex than the model we used to analyze the crime ring – the additional parameters suggested here can be determined through physical simulation of pathogen propagation, as well as through machine learning, using a large set of images as the training set. This is possible because there would be many fewer legal barriers to obtaining the volume of data required for a data-intensive approach.

5.4 Strengths and Weaknesses of the Model

The electric circuit model is effective in the case at hand because it is more expressive (and compatible with our approach philosophy) than a Markovian model, captures the important connectivities between nodes and topics, and throws everything else out. In other words, it is both expressive and simple.

Expressiveness is important because it prevents us from making unwarranted assumptions implicitly. For example, in a Markovian model, the probabilities of moving from any node to a subsequent node sum to 1 – such a model does not differentiate between a person making one suspicious and one unsuspicious phone calls, or ten of each, but does differentiate these from a person making three suspicious and ten unsuspicious calls (the third person would seem the least suspicious).

Simplicity is important because it prevents bloat in the number of parameters, which in turn reduces the risk of overfitting the data through a coincidental choice of parameters. We cut out direct node-to-node connections because their inclusion would result in three additional parameters, and thus, three additional assumptions. By considering the effect of adding a message between two nodes in the bipartite case, as well as empirical evidence from k-means clustering, it was clear that the benefit to our model of including node-to-node direct connections was outweighed by the risk of overfitting.

Our final model used only one single effective parameter: the ratio of α/β (that is, the relative weight of sent and received messages), which we set to 1.

There is also implicit simplicity in the circuit model due to its solvability via a linear system in Euclidean Space. This allowed us to meaningfully co-validate our approach against results from k-means clustering, for which we used the Euclidean Norm to form the distance measure.

Finally, the model is easily extensible to include additional information. Almost any additional information can be classified as affecting node-to-node, node-to-topic, or topic-to-topic conductivity, so the model provides a clear approach to include new data. We discuss some possible uses of such data above in the proposed application to a biological model, as well as to additional message data as described in Section 7.

One weakness of our model, particularly in the bipartite case, is that the benefits of free-flowing current do not help us analyze cases in which a node only sends a few messages, and these messages happen to be linked to suspicious topics. This fixes these nodes to a voltage of 1, placing them at the top of the ranking. This problem is mitigated for more well-connected nodes. However, although such nodes do not necessarily display a high volume of suspicious traffic (and thus, might not deserve the maximum possible voltage), it is still extremely suspicious if a node only interacts with suspicious topics. Cases such as these are also better analyzed using full-text message contents – for example, one would ideally want to distinguish between a person who is supportive of or against the central topic in a message, especially if the single message plays a large role in their final ranking.

6 Results of Model

In this section, we will present results of applying our circuit model to the given data (message transactions, known conspirators, suspicious topics, etc.) and answer impending questions such as involvement of the managers. We also consider changes to the data, such as more or fewer known guilty people and/or topics, and observe the consequences on our results. Finally, we provide a means of setting the cut off that splits the employees into two categories: investigate or don't investigate.

6.1 Immediate Results

In Table 3 we list the results from our model for two cases: the original conditions given and taking Chris and Topic 1 to be guilty. This listing shows top 15 employees most likely to be guilty, ranked by guilt level (voltage), along with associated voltages. A complete listing of all 82 employees and their associated guilt levels can be found in Appendix A.

We can note an interesting feature of the results (ignoring the Chris and Topic 1 being guilty case for now). Dayi and Sheng have a voltage of 1, which signifies that they only communicated about guilty topics. For example, Dayi only participated in a single message, which was with Dolores about Topic 7 (considered to be conspiratorial). While this may not be enough evidence to prosecute Dayi, the fact that his only message was conspiratorial in nature is certainly enough to warrant further investigation. We keep in mind that our main goal is to prioritize the employees by likelihood of being part of the conspiracy, and someone like Dayi who partook in only conspiratorial messages should be a top priority.

6.1.1 Involvement of Managers

First, let us evaluate the involvement of senior managers Jerome, Dolores, and Gretchen in the conspiracy. Not counting the 7 known conspirators, Dolores ranks 10th out of 75 in terms of guiltiness (as measured in our model). Given that she is a senior manager, the implications of her involvement in the conspiracy are greater, so ranking in the top 10 should warrant further investigation.

The situation with Jerome and Gretchen is a bit more intriguing, because there are two nodes named "Jerome" and two nodes named "Gretchen" in the data supplied. As we explained in the Section 4, we believe (through clustering results) that there are two people named Jerome and two people named Gretchen, so that these four nodes represent four unique people. That being said, one of these Jeromes ranks 5th out of 75 in guiltiness while the other ranks 43 out of 75. The guiltier Jerome participated in a total of 8 messages while the less guilty Jerome participated in 20. Since senior managers tend to receive and send more messages than lower ranking employees, our best hypothesis is that the manager Jerome was the less guilty one. He ranked in the bottom half in terms of guiltiness, so unless over half of the entire company was involved in the conspiracy, this manager is likely to be innocent.

Chris Innocent, Topic 1 unknown			***Chris and Topic 1 Guilty***		
Node #	**Name**	**Guilt (V)**	**Node #**	**Name**	**Guilt (V)**
51	Dayi	1	51	Dayi	1
57	Sheng	1	56	Cha	1
81	Seeni	0.956521739	57	Sheng	1
80	Fanti	0.877192982	81	Seeni	0.969465649
16	Jerome	0.859649123	16	Jerome	0.923076923
56	Cha	0.842105263	80	Fanti	0.915789474
79	Phille	0.839285714	10	Dolores	0.896103896
28	Dwight	0.836734694	13	Marion	0.89261745
33	Kim	0.831168831	28	Dwight	0.888111888
10	Dolores	0.828571429	4	Gretchen	0.883116883
75	Bariol	0.826086957	75	Bariol	0.882352941
4	Gretchen	0.825892857	41	Donald	0.87628866
13	Marion	0.81981982	63	Quan	0.875621891
60	Lars	0.804878049	79	Phille	0.875621891
47	Christina	0.798816568	69	Han	0.87477314

Table 3: The 15 guiltiest employees, excluding those who are given to be guilty, are listed in order of guilt, with guilt level to the right of an individual's name and node identification number to the left. The left half shows results from the original given data (Chris is known to be innocent and the nature of Topic 1 is undetermined), and the right half shows how the results change when Chris and Topic 1 are taken to be guilty.

Similarly, one Gretchen ranked 12th out of 75 while the other ranked 37th. The guiltier Gretchen participated in 14 messages while the less guilty participated in 23. Using the same reasoning as above, our best inference is that the manager Gretchen is not likely to be guilty.

Note that the inferences made from our limited data can easily be verified or corrected with supplementary information that would be easy to obtain in real life. For example a company directory or employees would tell us whether or not there are two Jeromes and two Gretchens. Similarly, if there are two Jeromes or Gretchens, looking at the email address or phone number associated with a message would indicate whether it was the manager or not.

6.1.2 Effects of Additional Known Conspirators or Topics

From the results in Table 3, we can observe what happens when Chris, who was previously given to be innocent, is taken to be guilty, and Topic 1 is taken to be conspiratorial. Note that the 15 guiltiest people, other than those known to be guilty beforehand, are largely conserved. 12 employees appear in the top 15 both before and after the conspiratorial natures of Chris and Topic 1 are modified. This instills a degree of confidence in our model because if one out of eight known criminals is mistakenly believed to be innocent and a conspiratorial topic is not deciphered, the model still works in identifying the top suspects.

However, we also note that the order of the suspects gets shuffled around, and that certain individuals— Kim, Lars, Christina, Donald, Quan, and Han—appear in the top 15 of one list but not the other. This is expected because a person who communicates a lot with Chris and talks about Topic 1 a lot will be considered guiltier if Chris and Topic 1 are taken to be guilty.

In short, our model demonstrates a sensitivity to given conditions, as it should, but remains robust enough to identify suspects when one out of eight people and one of four topics are classified incorrectly.

6.2 Robustness of Results

6.2.1 Leave-One-Out Testing

Here we consider the effects of treating a single person or topic given to be guilty as an unknown (voltage is not fixed to 1 V). We run our model for all 10 cases of leaving out a given (7 known conspirators and 3 suspicious topics). The results for leaving out known conspirators is shown in Table 4, while those for leaving out known conspiratorial topics is shown in Table 5. The purpose of this test is two-fold: it shows robustness of the model and prevents over-reliance on guilt by association.

The idea is that a person should not be accused solely on the premise of their association with a particular person or topic. For example, if my boss was known to be guilty and

None		***Harvey***		***Yao***		***Jean***	
Name	**Guilt (V)**	**Name**	**Guilt (V)**	**Name**	**Guilt (V)**	**Name**	**Guilt (V)**
Dayi	1	Dayi	1	Dayi	1	Dayi	1
Sheng	1	Sheng	1	Sheng	1	Sheng	1
Seeni	0.956521739	Seeni	0.955555556	Seeni	0.956043956	Seeni	0.95483871
Fanti	0.877192982	Fanti	0.87394958	Yao	0.890625	Fanti	0.871559633
Jerome	0.859649123	Jerome	0.85620915	Fanti	0.875816993	Jerome	0.853211009
Cha	0.842105263	Cha	0.84	Jerome	0.858108108	Cha	0.838383838
Phille	0.839285714	Phille	0.833333333	Cha	0.840909091	Phille	0.834782609
Dwight	0.836734694	Dwight	0.832946636	Phille	0.837837838	Dwight	0.830188679
Kim	0.831168831	Kim	0.828571429	Dwight	0.835051546	Kim	0.823170732
Dolores	0.828571429	Dolores	0.825	Kim	0.828571429	Dolores	0.821428571

Paul		***Alex***		***Elsie***		***Ulf***	
Name	**Guilt (V)**	**Name**	**Guilt (V)**	**Name**	**Guilt (V)**	**Name**	**Guilt (V)**
Dayi	1	Dayi	1	Dayi	1	Dayi	1
Sheng	1	Sheng	1	Sheng	1	Sheng	1
Seeni	0.95505618	Seeni	0.955882353	Seeni	0.954314721	Seeni	0.956521739
Fanti	0.871287129	Alex	0.890909091	Fanti	0.871794872	Ulf	0.948275862
Jerome	0.852941176	Fanti	0.875486381	Jerome	0.853658537	Fanti	0.876923077
Cha	0.837209302	Jerome	0.857142857	Cha	0.836734694	Jerome	0.85915493
Phille	0.83351469	Phille	0.837837838	Phille	0.835051546	Cha	0.841463415
Dwight	0.830188679	Cha	0.835820896	Dwight	0.829787234	Phille	0.83908046
Kim	0.825	Dwight	0.834482759	Kim	0.826923077	Dwight	0.836363636
Dolores	0.820512821	Kim	0.83	Dolores	0.822222222	Kim	0.830769231

Table 4: The 10 guiltiest people (not taken to be guilty) are shown below for the cases of leaving out each known conspirator as an unknown. The person left out is shown in the top row.

None		***Topic 7***		***Topic 11***		***Topic 13***	
Name	**Guilt (V)**	**Name**	**Guilt (V)**	**Name**	**Guilt (V)**	**Name**	**Guilt (V)**
Dayi	1	Sheng	1	Dayi	1	Dayi	1
Sheng	1	Seeni	0.850340136	Sheng	1	Seeni	0.915662651
Seeni	0.956521739	Cha	0.827586207	Seeni	0.913333333	Fanti	0.864864865
Fanti	0.877192982	Jerome	0.824742268	Cha	0.826923077	Phille	0.829545455
Jerome	0.859649123	Phille	0.823529412	Fanti	0.802631579	Jerome	0.826666667
Cha	0.842105263	Kim	0.817647059	Dolores	0.791666667	Bariol	0.8125
Phille	0.839285714	Bariol	0.80794702	Dwight	0.791666667	Dwight	0.800475059
Dwight	0.836734694	Fanti	0.807692308	Lars	0.779661017	Dolores	0.800420168
Kim	0.831168831	Dayi	0.80104712	Jerome	0.777027027	Gretchen	0.796296296
Dolores	0.828571429	Gretchen	0.790697674	Donald	0.766233766	Lars	0.79047619

Table 5: The 10 guiltiest people (not taken to be guilty) are shown below for the cases of leaving out each known conspiratorial topic as an unknown. The topic left out is shown in the top row.

Not Merged		***Merged***	
Name	**Guilt (V)**	**Name**	**Guilt (V)**
Dayi	1	Dayi	1
Sheng	1	Sheng	1
Seeni	0.956521739	Seeni	0.958490566
Fanti	0.877192982	Fanti	0.88372093
Jerome	0.859649123	Phille	0.848484848
Cha	0.842105263	Cha	0.847058824
Phille	0.839285714	Dwight	0.844262295
Dwight	0.836734694	Kim	0.842105263
Kim	0.831168831	Dolores	0.837606838
Dolores	0.828571429	Bariol	0.833641405

Table 6: The left side of the table shows results when all names that occur twice are treated as unique individuals, except for Beth. The right side shows results for when all twice-occuring names are merged into one individual.

we frequently exchanged messages, we would have a strengthened connection in the circuit model. However, this does not necessarily imply that I am guilty, so we want to be able to treat my boss as an unknown and then evaluate my guiltiness in that circumstance.

We can see from the results in Tables 4 and 5 that most of our results pass the "leave-one-out" test. We note that 6 people rank in the top 10 for each of the ten leave-one-out situations and can thus be investigated without complaint against grounds of guilt by association. This also demonstrates robustness of the model in the sense that if less information existed on who (or what topic) was known to be guilty, we could still identify the most likely suspects.

Finally we note that our model is consistent with the given information. For example, when Yao, Alex, and Ulf are treated as unknowns, they still place within the 10 guiltiest, signifying that they could be caught even if they weren't know to be guilty. While the other four known conspirators don't rank within the top 10 when left as unknowns, they do still rank high. This lends some credibility to our model, as the results check against known data.

6.2.2 Merging Repeated Names

Next we demonstrate that the ambiguity of repeated names in the data do not seriously affect the results of our model. Table 6 compares the results of treating each node as an individual (except the two Beths)—as we have been doing all along—and treating each name as one individual (we assume the that two "Jerome" nodes refer to the same person). As we can see from the table, the difference between these two treatments is small. The ordering is largely preserved, and the only people who differ among the top 10 are Jerome (who only appears in the unmerged set) and Bariol (who only appears in the merged set). Note that Bariol is actually 11th in the unmerged list, so his ranking is almost exactly the same. Jerome, on the other hand, drops to 20th in the merged list, which is intermediate between the guiltier Jerome (5th) and the less guilty Jerome (43rd). Again, this shows that our model is largely resilient against the ambiguities present in the data, and the sensible ranking of the combined Jerome is further demonstration that our model behaves in accordance with expectations.

6.3 Determining Potential Cutoff

The circuit model excels at prioritizing the employees for investigation, but a weakness is that the cutoff voltage for distinguishing between "maybe guilty" and "most likely not guilty" is unclear. One way of getting around this setback is to determine the cutoff based on reasonable assumptions about the average number of people involved in conspiracies. For example, in 2009, 592 individuals were investigated for financial crimes relating to corporate fraud, and 156 individuals were convicted [3]. Some of the major cases that year involved the companies Petters Company, Inc., with 3,200 employees, and Credit Suisse, with 50,000 employees. Even in the extreme case that all investigated individuals were from the smaller Petters Company, they would still only account for 18.5% of the company. Thus, drawing the cutoff at the 20% line—17 out of the 82 employees in our case—would be conservative (with respect to including as many potential criminals as possible) by historical standards. Since 7 of these 17 employees to be investigated are already given to us, this amounts to looking at the next 10 highest ranking (by determined guilt) individuals, which is what we have been doing in the sections above.

7 Future work

In this section, we outline possible avenues for improving our model upon the availability of message full-text data.

7.1 Semantic Network Analysis

Our model currently does not form any direct connections between topics based on how closely related they are. This is largely because the small volume of text given as topic descriptions is too small for effective Semantic Network Analysis, and cannot be effectively filtered using Latent Semantic Indexing (which performs dimensionality reduction through singular value decomposition to search for common strands of co-occuring words). An experimental attempt at using WordNet (a database mapping lexical and conceptual-semantic relations between words) and breadth-first (as a means of shortest-path finding) to determine the distance between related words was also unsuccessful due to adversely large distances between many words.

The existence of message full-text contents would allow us to compile such message contents into documents representing each topic, and use Latent Semantic Indexing on the whole corpus to identify sets of co-occurring words. This would allow us to de-noise the topic documents and reduce their dimensionality. An inner product of the resulting document vectors (post-normalization) would provide a measure of topic similarity which we could use as inter-topic conductances (higher similarity yields greater conductance). It would also help make further Semantic Network Analysis both tractable and meaningful, which could afford a richer measure of topic similarity than the document vector inner products alone.

7.2 Text Analysis

A key aspect of text analysis that would likely improve our model is sentiment analysis. Several of the fifteen given topics revolve around specific nodes, but because these topics all revolved around controversy, it was difficult for us to glean node-to-node information from topic associations alone. Sentiment analysis of the full-text contents of messages can give us an understanding of the feelings that nodes harbor for eachother, which would result in a compelling measure of direct node-to-node conductivities, which we currently do not use. This is analogous to using effective proximity as node-to-node conductance in the sample biological problem application given in the “Applicability to Generic Agent-Ranking Problems” subsection of Section 5.

Apart from enriching our model, introducing meaningful node-to-node conductances could help reduce the sensitivity of the voltages of nodes who don’t communicate often, by using a network measure that doesn’t simply count the communications and topics themselves.

8 Recommendation to the DA

Working under the original assumptions of 7 known conspirators and 8 known non-conspirators, our model suggests that the 10 next most likely candidates for being conspirators are, in order: Dayi, Sheng, Seeni, Fanti, Jerome (non-manager), Cha, Phille, Dwight, Kim, and Delores. Based on historical precedence, we expect that from a pool of 82 suspects, less than 17 employees on average will be involved with the conspiracy. We therefore recommend investigating these ten workers, who along with the 7 known conspirators make 17. However, given the relatively small company size, it could be the case that there are actually more than 17 conspirators, or that a few of the most likely candidates are actually innocent, in which case it would be prudent to continue investigating subjects in the order listed within the appendix. The extent to which you continue to investigate would depend on the resources available.

References

[1] Linguistic inquiry and word count. `http://liwc.net/index.php`.

[2] Peter G. Doyle and J. Laurie Snell. Random walks and electric networks. January 2000.

[3] FBI. Financial crimes report. http://www.fbi.gov/stats-services/publications/financial-crimes-report-2009, 2009.

[4] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[5] Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. Lying words: Predicting deception from linguistic styles. 2003.

[6] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

A Complete Listing of Results

All 82 employees of the company are listed below in order from most likely to be guilty to least likely to be guilty. The guilt levels / voltages determined by our model are shown to the right of the corresponding employee, along with the node identification number to the left. The italicized entries at the top and bottom of the table are those people given to be guilty ($V = 1$) or innocent ($V = 0$). These results are for the original given data (i.e. Chris is innocent, Topic 1 is not known to be guilty, etc.).

Node #	**Name**	**Guilt (V)**
7	*Elsie*	*1*
18	*Jean*	*1*
21	*Alex*	*1*
38	*Paul*	*1*
44	*Harvey*	*1*
49	*Ulf*	*1*
67	*Yao*	*1*
51	Dayi	1
57	Sheng	1
81	Seeni	0.956521739
80	Fanti	0.877192982
16	Jerome	0.859649123
56	Cha	0.842105263
79	Phille	0.839285714
28	Dwight	0.836734694
33	Kim	0.831168831
10	Dolores	0.828571429
75	Bariol	0.826086957
4	Gretchen	0.825892857
13	Marion	0.81981982
60	Lars	0.804878049
47	Christina	0.798816568
72	Andra	0.794871795
41	Donald	0.793650794
50	William	0.788235294
6	Patrick	0.782608696
46	Louis	0.78
22	Eric	0.778947368
17	Neal	0.778169014
19	Kristine	0.774011299
3	Sherri	0.773809524
38	Beth	0.773584906
36	Priscilla	0.772357724

29	Wayne	0.770833333
23	Wesley	0.769230769
15	Julia	0.766990291
39	Erica	0.764285714
20	Crystal	0.762295082
69	Han	0.762295082
82	Reni	0.76
8	Hazel	0.75739645
27	Marcia	0.756302521
42	Katherine	0.753424658
32	Gretchen	0.74978355
40	Douglas	0.748427673
35	Shelley	0.747663551
11	Francis	0.746478873
30	Stephanie	0.745762712
31	Neal	0.745762712
34	Jerome	0.743119266
45	Lois	0.742268041
5	Karen	0.74137931
44	Patricia	0.74137931
37	Elsie	0.739361702
9	Malcolm	0.735849057
53	Chara	0.729166667
76	Cole	0.725806452
73	Carina	0.720430108
77	Gerry	0.72
24	Franklin	0.719512195
52	Vind	0.716981132
12	Sandy	0.714932127
1	Kristina	0.707317073
55	Olina	0.704225352
26	Marian	0.703703704
71	Cory	0.697674419
25	Claire	0.694267516
63	Quan	0.68115942
58	Lao	0.678362573
59	Darol	0.678362573
66	Melia	0.676767677
70	Hark	0.673469388
61	Le	0.646825397
62	Mai	0.64556962
0	*Chris*	*0*

2	Paige	0
48	Darlene	0
64	Tran	0
65	Jia	0
68	Ellin	0
74	Gard	0
78	Este	0