NETWORK ANALYSIS WITH THE ENRON EMAIL CORPUS

J.S. HARDIN, G. SARKIS, AND P.C. URC

ABSTRACT. We use the Enron email corpus to study relationships in a network by applying six different measures of centrality. Our results came out of an in-semester undergraduate research seminar. The Enron corpus is well suited to statistical analyses at all levels of undergraduate education. Through this note’s focus on centrality, students can explore the dependence of statistical models on initial assumptions and the interplay between centrality measures and hierarchical ranking, and they can use completed studies as springboards for future research. The Enron corpus also presents opportunities for research into many other areas of analysis, including social networks, clustering, and natural language processing.

1. Introduction

One of the most infamous corporate scandals of the past few decades curiously left in its wake one of the most valuable publicly-available datasets. In late 2001, the Enron Corporation’s accounting obfuscation and fraud quickly led to the bankruptcy of the large energy company. The Federal Energy Regulatory Commission subpoenaed all of Enron’s email records as part of the ensuing investigation. Over the following two years, the commission released, unreleased, and rereleased the email corpus to the public after deleting emails that contained personal information like social security numbers. The Enron corpus contains an email spectrum from weekend vacation planning to political strategy talking points, and it remains the only large example of real world email datasets available for research. See [FER, 2013] for the Federal Energy Regulatory Commission’s website on the Enron investigation, [Dat, 2003] for the final order releasing the data to the public, and [McLean and Elkind, 2013] for a popular account of the Enron scandal.

In this note, we discuss six centrality measures of the Enron corpus and suggest how they can be used in undergraduate education and research. Our results and methods came out of an undergraduate research circle at Pomona College that we oversaw during the spring semester of 2014. The research circle consisted of four students whose interests and initiative determined the research questions and research direction, and two math/stats faculty members who provided general and technical guidance.

The Enron corpus is in many ways an ideal dataset for statistical pedagogy. Its origins are unusual and engaging for students who are interested in real-world data and recent American econo-cultural history; a sizeable literature already exists on the corpus, so that students need not start the conversation and investigation at square zero; social networks are accessible, especially in this post-Facebook era, yet they motivate current and active research problems; centrality measures are intuitive and mathematically nontrivial; and the discussion presented below may be used for stand-alone research modules in an undergraduate statistics class or may serve as a starting point for a more intensive research project.

1.1. Literature. Research into the corpus has been prolific and wide ranging. We present here a selection of the publications about Enron to highlight the range of research that the corpus has spurred, and to suggest possible further directions as well.

See [Shetty and Adibi, 2004] for a technical report describing a MySql database of the corpus, [Wang et al., 2014] for anomaly detection in a dynamic network, [Diesner et al., 2005] for a social network analysis that focused on changes in behavior during the scandal period, [Deitrick et al., 2012] for a neural networks model predicting the gender of an emailer based on the email stream, [Peterson et al., 2011] for measures of formality in the email correspondence, [Chapanond et al., 2005] for a graph-theoretic and spectral analysis that overlaps with many of the topics of interest in our note, [Martin et al., 2005] for detection of abnormal email

Date: October 13, 2014.
P.C. URC stands for the Pomona College Undergraduate Research Circle whose members for this project were Timothy Kaye, David Khatami, Daniel Metz, and Emily Proulx.
activity in outgoing messages, [Zhou et al., 2006] for a probabilistic approach to community detection, and [Zhou et al., 2007] for data cleaning with focus on email aliases.

1.2. Research questions. We are interested in the social network that is defined by the emails. In particular, we investigate what kinds of information about the relative importance of the Enron employees can be read from the graph whose vertices are the employees and whose edges represent email correspondence.

There is no reason to believe that the kinds of importance ranking induced by an email connectivity graph reflects the managerial structure of the corporation itself. Indeed, rankings based on email networks point to an overlay of the activity of an individual emailer and the subnetwork of contacts that emailer has. As such, centrality measures based on an email network may help gauge the functional importance of various employees, as opposed to (or in conjunction with) their managerial importance; and different centrality measures are more adept at spotting different functionalities. By addressing the research question about importance in the email corpus, we also survey centrality measures and clustering methods and analyze of some of their strengths and limitations.

1.3. Network Analysis. Networks are ubiquitous in the internet age, underlying much of virtual (and real) life from social webs to recommender systems, and from epidemiological spread to linguistic evolution. They are used widely as tools of research in sociology [Sutton et al., 2014] (patterns of Tweets during a natural disaster), biology [Pinter-Wollman et al., 2011] (coordinated behavior of harvester ants), genetics [Zhang and Horvath, 2005] (co-expressed gene groups in brain cancer), and economics [Stephen and Toubia, 2010] (economic value of a social network in a large online marketplace) to study the behavior of individuals and of systems.

The mathematical model of a network can be used to identify any of a variety of characteristics of the system. For example, the key question about the network might be the centrality—which node is most important or influential in the system. Alternatively, it might be important to understand which subsets of nodes are connected. Other metrics include network closure (how complete the network is—whether a node which is connected to a second node is also connected to the second node’s connections); distance (the number of degrees of separation between any two nodes); and cliques (if every node is connected to every other node).

Using the Enron corpus, we consider two nodes to be connected if their corresponding employees have corresponded by email. We weight the connection, or the edge between the two nodes, by the number of emails sent. Additionally, we use directionality to separately analyze emails sent or emails received, when appropriate. We give examples below of some of the network information resulting from the various choices of adjacency metrics. Our analysis here focuses on finding the central nodes in the email network. Additionally, we provide a brief analysis of the group membership of the most connected cliques, found by hierarchical clustering.

1.4. In the Classroom. We consider the topics to be upper level undergraduate techniques which could easily be taught in a multivariate statistics course, a machine learning computer science course, or a data science course. Additionally, network analysis or clustering could easily be added as a topic to a course on statistical applications.

We also see a place for the topics in math classes that look for applications to their methods. In an analysis class that covers metric spaces, networks provide an interesting field of play. In a linear algebra class, eigenvector centrality can make the mathematical theory come alive. We detail some of the specific ways our examples can be used in different statistics and mathematics courses in Section 4 below.

The use of recent and meaningful data improves the classroom experience in terms of both engaging students and solidifying their technical knowledge. It has been our experience that students are able to engage more thoughtfully with statistical methodology when they are interested in the research question at hand—an interest that is usually concurrent with providing intriguing data. Indeed, in our research circle, the students were given free range to choose both the data set to work with and the analysis method to apply for our semester long research project. They unanimously chose to work with the Enron corpus and apply network analysis to the email counts.
1.5. **Dataset Story: Cleanup and Processing.** The narrative aspect of many datasets in both pedagogy and research includes a major data-collection component. Even in classroom examples where the data, or a summary thereof, is given to the students, there often exists a contextual story about how and why the data might have been collected for the immediate purpose of the statistical analysis. The Enron corpus, on the other hand, is for all intents and purposes an accidental, incidental dataset. This presents an invaluable opportunity to discuss real-world data issues that do not often come up in the classroom. Specifically, real-world data is often dirtier and less cooperative than experimental data. It is not structured with a specific goal in mind—it is what it is. Therefore, getting it to the tidy stage where analysis may be conducted and meaning may be extracted involves several assumptive and simplifying decisions that require thoughtful analysis before the fact.

For our project, we used the dataset available at [https://s3.amazonaws.com/metanautix/enron/enron-mail_20110402_csv.tgz](https://s3.amazonaws.com/metanautix/enron/enron-mail_20110402_csv.tgz), whose emails were organized into 150 mailboxes labeled by employee name; the emails in a mailbox were not necessarily sent by that person. Additionally, some employees with similar names were binned into the same mailbox, while others had their messages split among two mailboxes. In order to circumvent such potential binning errors, we ignored the folder designation and instead extracted only From, To, and CC fields of each email message. While only one employee may appear in either the From or To fields, an arbitrary number may appear in the CC field. We considered only senders and recipients with email addresses that have an enron.com domain name. To distinguish between the individuals, we relied on six standard aliases used at Enron (see [Zhou et al., 2007](#) for instance). The result was 156 employees whose email communication we considered, and from which we constructed an adjacency matrix for the weighted directed graph of Enron employees, as follows.

![Figure 1](https://s3.amazonaws.com/metanautix/enron/enron-mail_20110402_csv.tgz)

**Figure 1.** Each $(i, j)^{th}$ dot represents a binary indicator that an email was sent from person $i$ to person $j$.

Let $E_{ij}$ be the set of emails for which Enron employee $i$ appears in the From field and employee $j$ appears in the To field. Let $C_{ij}$ be the set of emails for which Enron employee $i$ appears in the From field and employee $j$ appears in the CC field. For each $c \in C_{ij}$, let $n_c$ be the number of names that appear in the CC
field of $c$. Define the $156 \times 156$ weighted adjacency $M$ as:

$$(1) \quad m_{ij} = |E_{ij}| + \sum_{c \in C_{ij}} \frac{1}{\sqrt{1+n_c}}$$

Thus, for the weighting of the edge in the directed graph from employee $i$ to employee $j$, each email sent from $i$ to $j$ contributed 1, and each email $c$ sent from $i$ on which $j$ was cc-ed contributed $1/\sqrt{1+n_c}$.

The matrix $M$ is the weighted adjacency matrix of a directed graph. To consider the undirected graph, we defined the matrix $U = M + M^T - D$, where $D$ is the diagonal matrix $d_{ii} = 2m_{ii}$. In other words, for the undirected graph, we did not incorporate information from emails that employees cc-ed themselves on.

To highlight the importance of data cleanup decisions, even if that is tangential to the focus of this paper or a class, students can discuss the multitude of ways to represent the Enron email network, and the potential consequences to the analysis of each decision or assumption made along the way. For instance, are there employee-specific parameters that can be computed without constructing the entire network? Also, students can discuss different weightings for the matrix $M$. What if being emailed directly and being cc-ed counted equally? Is there a way to incorporate the importance of a message in the weighting, say by a blunt measure like the length of the email, or by a more sophisticated textual analysis?

## 2. Centrality and Rank

A measure of centrality on a graph aims to assign a ranking or magnitude to each node that captures the relative importance of that node in the context of the graph’s structure. We are interested in measuring the importance of each employee based on the number of emails sent or received, as aggregated in the dataset we extracted from the Enron corpus and summarized in the matrices $M$ and $U$. Recall that $M$ is the weighted $156 \times 156$ adjacency matrix of our directed graph, and $U$ is the corresponding weighted matrix of the undirected graph that does not distinguish between sent and received emails.

We investigate six measures of importance within the Enron employee email network: degree/strength, eigenvector centrality for received emails, eigenvector centrality for sent emails, closeness, betweenness, and topological overlay. We give an overview of these measures below, including mathematical definitions and intuition. For each of these measures, it may be of interest for students to generate examples of nodes in a network that rank high or low in centrality. In Section 3 below, we also suggest how some of these measures may be incorporated into statistics classes of various levels.

### 2.1. Degree and strength

The degree $\delta_i$ of employee $i$ is defined to equal the total number of employees to whom $i$ sent or received emails. Thus, if we define

$$\delta_{ij} = \begin{cases} 1 & \text{if } u_{ij} \neq 0 \\ 0 & \text{if } u_{ij} = 0 \end{cases}$$

then $\delta_i = \sum_j \delta_{ij}$. We did not distinguish between whether $j$ appeared in the To or CC field. The degree is a measure of the size of $i$'s immediate network. The more different people $i$ emails, directly or by cc, the greater $i$'s degree. For instance, an employee who forwards a single announcement to everybody in the company can achieve maximal degree.

The strength $\sigma_i$ of employee $i$ is defined to equal the total number of emails that $i$ sent or received. In our implementation, we computed $\sigma_i = \sum_j u_{ij}$. Strength is also a size measure, but of the volume of $i$'s correspondence instead of the extent of $i$'s network. The more emails $i$ sends, the greater $i$'s strength. An employee can rank higher on the strength centrality by simply activating an automated reply.

For strength and degree, we incorporated our cc-weighting system so that our results might be compared to other measures using the matrix $U$ of the weighted, undirected graph. It may be of interest to explore how the results might change by using a binary (unweighted) matrix instead.

While each of strength and degree is a blunt measure of importance on its own, the two can be effectively combined. Following [Opsahl and Skvoretz, 2010], let $\alpha \geq 0$ be a “tuning parameter,” and define the $\alpha$-combined measure $\kappa_i(\alpha) = \delta_i^\alpha \sigma_i^{1-\alpha}$. For our implementation, we chose $\alpha = 0.5$ to give equal weight to degree and strength. Different values of the tuning parameter obviously result in different measures of rank, and provide an opportunity for further exploration. For example, if $\alpha > 1$, then $\kappa_i(\alpha)$ is penalized whenever...
i’s strength $\sigma_i$ is large—that is, when $i$ sends out a large volume of emails. Thus, large values of $\alpha$ may be used to scale the breadth of an employee’s network by the volume of that employee’s correspondence.

2.2. Eigenvector Centrality. Suppose the importance $x_i$ of employee $i$ is accumulated from the importances $x_j$ as $j$ ranges over all employees that $i$ emails. Suppose further that employee $j$ contributes to $x_i$ in proportion to the connectedness from $i$ to $j$ as measured by $m_{ij}$. That is,

$$x_i = \frac{1}{\lambda} \sum_j m_{ij} x_j$$

where $\frac{1}{\lambda}$ is some proportionality constant.

We can summarize these relationships with the matrix equation $M\vec{x} = \lambda\vec{x}$, where $\vec{x} = (x_1 \cdots x_{156})^T$. In other words, $\vec{x}$ is an eigenvector of $M$ with eigenvalue $\lambda$. While $M$ may have several different eigenvalues and eigenvectors, the Perron-Frobenius Theorem (see [Horn and Johnson, 1990] page 508) for instance guarantees that because $m_{ij} \geq 0$, then for some eigenvalue $\lambda$ with largest absolute value, there exists an eigenvector $\vec{x}$ whose entries are all nonnegative; that so-called dominant eigenvector provides the importance weights for the employees.

We also consider $M^T$, the transpose of $M$, in order to analyze importance based on received emails. In that case, we compute $M^T\vec{x} = \lambda\vec{x}$ for the eigenvector of importance weights corresponding to the eigenvalue $\lambda$ with the largest absolute value.

Such a measure of importance is called eigenvector centrality. While it is commonly used with binary or stochastic matrices, its premise applies to any matrix with nonnegative entries. Arguably the most famous instance of eigenvector centrality is the first implementation of Google’s PageRank algorithm: webpages rank highly in Google’s search results if they are linked from other webpages of high rank. See [Austin, 2016] for a fun illustration of the algorithm suitable for a linear algebra class, and [Brin and Page, 1998] for the original paper by Google founders Sergey Brin and Lawrence Page.

An employee ranks highly in eigenvector centrality (respectively, transpose-eigenvector centrality) if that employee sends emails to (respectively, receives emails from) many other highly-ranked employees. Students can generate and discuss examples of company structure based on whether these two eigenvector centralities turn out to be highly correlated or uncorrelated: what kinds of employees might rank highly in eigenvector centrality of emails sent, but not in the transpose of emails received?

2.3. Closeness. Given a pair of employees $i$ and $j$, a path from $i$ to $j$ is defined to be a sequence $i_0 = i, i_1, \cdots, i_r = j$ such that $m_{i_{t-1}i_t} \neq 0$ for all $1 \leq t \leq r$. A path from $i$ to $j$ is called a shortest path if it minimizes the number of steps $r$ in the sequence, and the distance from $i$ to $j$, denoted $d(i,j)$, is the number of steps in such a shortest path from $i$ to $j$. The closeness of $i$ is then defined as

$$\gamma_i = \frac{1}{\sum_{j \neq i} d(i,j)}$$

If the graph has more than one connected component, the closeness of any node equals zero. Otherwise, the closeness of $i$ measures the speed or efficiency with which information spreads out from $i$ to the rest of the graph. Note that $\gamma_i$ is sometimes normalized by the number of nodes so that it measures the reciprocal of the average distance from $i$: $n/\sum_{j \neq i} d(i,j)$. However, the ranking of employees based on closeness is independent of such a normalization.

An employee has high closeness centrality if that employee’s correspondence reaches a large proportion of the network quickly. Thus closeness is a measure of the entire network’s structure in relation to a node. It may be interesting to discuss the robustness of the closeness centrality. For instance, can an employee rise in the rankings by sending one or two carefully chosen emails?

2.4. Betweenness. For betweenness, we consider the undirected network and its adjacency matrix $U$, and ask path-related questions similar to those in closeness measures. Given a pair of employees $j$ and $k$, an undirected path from $j$ to $k$ is defined to be a sequence $i_0 = j, i_1, \cdots, i_r = k$ such that $u_{i_{t-1}i_t} \neq 0$ for all $1 \leq t \leq r$; in particular, we do not take into account the from/to directedness of the graph of Enron employees. A path from $j$ to $k$ is called a shortest undirected path if it minimizes the number of steps $r$ in the sequence.
Since our paths are weighted, we make an adjustment (following [Newman, 2001, Opsahl and Skvoretz, 2010]) to the definition of “shortest” under the premise that a higher weight implies a faster connection, and hence, a shorter path. Let \(i_0, i_1, \cdots, i_r\) be a path from \(i_0\) to \(i_r\). Then the weighted length of this path is the reciprocal sum of the weights of the path’s edges:

\[
\sum_{1 \leq t \leq r} \frac{1}{w_{i_t-i_{t-1}}}
\]

Let \(\tau_{jk}\) be the number of shortest undirected paths from \(i\) to \(j\), and let \(\tau_{jk}(i)\) be the number of paths from among those in \(\tau_{jk}\) that pass through \(i\). Then the betweenness of \(i\) is given by

\[
\beta_i = \sum_{i \neq j \neq k} \frac{\tau_{jk}(i)}{\tau_{jk}}
\]

As such, the betweenness of \(i\) measures the importance of \(i\) as a central node in efficient communication between other nodes in the network. An employee has high betweenness centrality if that employee figures prominently in the email proximity of many pairs of colleagues.

The directed/undirected choices for closeness/betweenness, respectively, naturally generate discussion questions about the reasons for these choices and about how these measures might differ if alternate choices were made. As well, students can investigate different modifications for the shortest path in closeness to account for weighting.

2.5. Topological Overlap Matrix. Topological Overlap Matrix (TOM) extends the adjacency matrix from a measure of connectedness between two nodes only to a measure of connectedness between two nodes and the rest of the individuals in the dataset [Ravasz et al., 2002, Yip and Horvath, 2007]. Let \(u_{ij}\) be the measure of adjacency between nodes \(i\) and \(j\) as defined in Subsection 1.5. We define the matrix TOM as:

\[
TOM_{ij} = \frac{\sum_{l \neq i,j} u_{il}u_{lj} + u_{ij}}{\min\left(\sum_{l \neq i,j} u_{il}, \sum_{l \neq i,j} u_{ji}\right) + 1 - u_{ij}}
\]

This new adjacency matrix is then converted to a centrality measure by taking the row sum of the TOM. That is, the most central node will be the one who is most connected to the other nodes by way of third party connections. It is worth pointing out that TOM (like degree/strength centrality) directly accounts for the second degree connections, and so it will naturally produce different measures of importance than other centrality measures.

Topological overlap adjacency was originally designed to take as input unweighted networks, or binary matrices. In such a case, \(TOM_{ij}\) measures the proportion of overlap between \(i\)’s and \(j\)’s immediate neighbors. There are three natural avenues of TOM-discussion for students. First, one may ask if using the same formula for weighted networks, as we did, can result in misleadingly inflated entries in the TOM matrix—for instance, should the quadratic growth of \(\sum_{l \neq i,j} u_{il}u_{lj}\) be tempered with a square root? Second, students can consider the Generalized Topological Overlap Matrix (see [Yip and Horvath, 2007]) which measures the the network overlap of all neighbors within a fixed distance from any two nodes. And third, the row-sum measure of TOM centrality is essentially the degree centrality defined above; as such, the other measures of centrality discussed above can be applied to the TOM matrix as well.

2.6. Network Construction. Adjacency can be measured using number of emails exchanged, a binary indicator of email exchanged, sent email(s) only, received email(s) only, or TOM applied to any of the previous measures - among many other measures of adjacency. After establishing the measure of adjacency, network construction can be done using either hierarchical or partitioning methods. We briefly discuss hierarchical clustering, but note that the data could be used to evaluate and compare different distance-based clustering methods.

Hierarchical network construction (or clustering) starts by comparing all pairwise nodes and connecting the two nodes which are most similar. The second step connects the next two nodes which are most similar or connects a node to the already connected group using either average connectedness, minimum connectedness, or maximum connectedness. The construction happens by making one additional connection at each step until all nodes are connected into one group [Everitt et al., 2011]. The resulting dendrogram (see Figure
provide a graphical representation of the algorithm that is easy to visualize and interpret. In fact, the number of clusters can often be intuited by simply looking at the graph. However, hierarchical networks have the disadvantage that in building the network, once two nodes are connected, they remain connected. A possible project for students is to use permutation methods to evaluate the significance of the resulting clustering output. We demonstrate hierarchical clustering below, and we require at least 4 nodes to form a group.

3. Helpful Hints

Our results build on Kaye et al. [Kaye et al., 2014], a semester long research experience for a group of undergraduates at Pomona College.

3.1. Centrality. Using degree & strength, eigenvector centrality, TOM, betweenness, closeness, we rank the central importance of each of the individuals in the dataset.

| Degree & Strength | EVcent | EVcent Trans | Closeness | Betweenness | TOM |
|-------------------|--------|--------------|-----------|-------------|-----|
| Jeff Dasovich     | Tana Jones | Sara Shackleton | Robert Benson | Louise Kitchen | Jeff Dasovich |
| Mike Grigsby      | Sara Shackleton | Susan Bailey | Mike Grigsby | Susan Scott | Richard Shapiro |
| Tana Jones        | Stephanie Panus | Marie Heard | Louise Kitchen | Mike Grigsby | Steven J. Kean |
| Sara Shackleton   | Marie Heard | Tana Jones | Kevin M. Presto | Kenneth Lay | Mike Grigsby |
| Richard Shapiro   | Susan Bailey | Stephanie Panus | Susan Scott | Sally Beck | Tana Jones |
| Steven J. Kean    | Kay Mann | Elizabeth Sager | Scott Neal | Jeff Dasovich | Sara Shackleton |
| Susan Scott       | Louise Kitchen | Jason Williams | Barry Tychohilz | Kevin M. Presto | Mary Hain |
| Louise Kitchen    | Elizabeth Sager | Louise Kitchen | Greg Whalley | Chris Stokley | Marie Heard |
| Stephanie Panus   | Jason Williams | Jeffrey T. Hodge | Phillip K. Allen | Mary Hain | Stephanie Panus |
| Marie Heard       | Jeff Dasovich | Gerald Nemec | Jeff Dasovich | Kate Symes | Susan Scott |

Table 1. According to each of six different measures of centrality, we provide a ranked list of the individuals who are most central to the email corpus.

3.2. Network. Using the R package Weighted Gene Co-expression Network Analysis (WGCNA) [Langfelder and Horvath, 2008] we cluster the observations into a hierarchical dendrogram. WGCNA uses a hierarchical clustering algorithm to an agglomerative (building one step at a time from 156 groups until all individuals are in one group) algorithm to link individuals sequentially based on the number of emails exchanged. We used average-linkage to determine closeness to a group that has already been formed; that is, an individual will be added to a group if they are close, on average, to the members of the existing group. Additionally, we did not require that every individual be linked into a group. We require that the dissimilarity be no more than 0.9 for the adjacency matrix. (Recall that the adjacency score is determined by the number of emails sent and received, divided by the maximum adjacency score. The dissimilarity is one minus the adjacency.) We require the TOM dissimilarity to be no more than 0.95. Lastly, each group is required to have at least 4 members according to our analysis. The dissimilarity measure, linkage decision, and cutoff criteria are all parameters that can be adjusted in order to gain further insight into the data.

4. Further Directions

We presented above some suggested directions that students can take with discussion and research questions for each of the centrality measures. We add to these here with some suggestions for class-specific modules and further exploration.

4.1. Connections to Specific Courses.
4.1.1. Introductory Statistics. The analyses done in this note are not typically covered in Introductory Statistics. However, the data could be used to do descriptive statistics. For example, students could make boxplots across different Enron departments using either number of emails sent or number of emails received. One might be able to run an inferential (e.g., chi-square) test to see if lawyers sent more emails to other lawyers or to non-lawyers. Indeed, an interesting classroom discussion could be based on the data clearly not being a representative sample from a population; instead, the data might be thought of as a sample from a process of email sending by the 156 individuals measured.

4.1.2. Applied Statistics. The data and analyses provided seem most appropriate for an applied statistics course with an introductory prerequisite (e.g., computational statistics, multivariate analysis, or data science). The Enron data allow for a complete analysis of centrality metrics as well as a consideration of different network or clustering construction methods which are based on distances. We have provided R code for an initial analysis, but our work could easily be expanded to include additional centrality measures or other network and clustering construction methods.

4.1.3. Mathematical Multivariate Analysis or Linear Algebra. Principal component analysis is a mainstay of multivariate analysis classes, and increasingly, eigenvector centrality makes a late-semester appearance in linear algebra classes. We submit that eigenvector centrality is at least equally as appropriate for a class in multivariate analysis in addition to, or instead of, PCA. Both PCA and eigenvector centrality require some linear-algebraic sophistication and dexterity with eigentheory. However, eigenvector centrality can be more intuitive—as the importance formula $x_i = \frac{1}{X} \sum_{j} x_j$ is a straightforward linear transcription of the
Figure 3. Dendrograms representing hierarchical clustering with the symmetric adjacency matrix as well as the TOM construction based on the symmetric adjacency matrix. We group points based on similarity (0.9 for symmetric adjacency and 0.95 for TOM) as well as a minimum cluster size (4 individuals).

importance-voting assumption of the model—while it still includes sophisticated machinery like the Perron-Frobenius Theorem. On the other hand, the connection between eigenvectors of the covariance matrix and the principal axes of a best-fit ellipse can be obscure to the student upon first introduction.
4.2. Alternative Applications.

4.2.1. Correlation between centrality and company hierarchy. The managerial hierarchy of Enron is not reflected in the top ten employees as ranked by the centrality measures above. Indeed, of the main executives at the company, only two appear in the top ten: Kenneth Lay, the CEO and chairman, came in fourth on the betweenness scale, and Greg Whalley, the president, had the eighth highest closeness score. While some studies have attempted to reconstruct the company hierarchy from the email network—see for instance [Agarwal et al., 2012] for an attempted recovery of dominance relationship from among the employees with known dominance-subordinate hierarchy by simply using the degree centrality—we are not aware of any studies that carefully interpret the significance of high rank in centrality measures in the context of the company’s hierarchy.

4.2.2. Gender and department. One of the interesting outcomes of our rankings is that the top eight scorers in eigenvector centrality were women. Also, most of the top ten eigenscorers were lawyers. There exists published studies that discuss email changes over time by department (see for instance [Diesner et al., 2005]), though these do not correlate the departments to the employees’ centralities. And while the Enron corpus has been used to study gender-related questions (like predicting gender from the email stream in [Deitrick et al., 2012]), we are not aware of centrality analyses of the Enron corpus with gender as a variable.

4.2.3. Generalized TOM and other centrality measures applied to TOM. As mentioned above, TOM can be generalized to m-step neighborhoods to measure agreement between nodes with respect to multiple steps of adjacency [Yip and Horvath, 2007]. Generalized TOM defines paths of length m to define adjacency between nodes. Additionally, a straightforward extension of TOM is to use other measures of adjacency (e.g., the binary measure of emails sent between two nodes) within the TOM metric. Alternatively, applying centrality measure like eigencentrality or closeness to the TOM matrix instead of the graph adjacency matrix may result in deeper centrality measures that better take into account overall network connectedness.

4.2.4. Tuning parameter. As described above, the degree $\delta_i$ and strength $\sigma_i$ of an employee $i$ can be effectively combined with a tuning parameter $\alpha$ to define the new centrality measure $\kappa_i(\alpha) = \delta_i^\alpha \sigma_i^{1-\alpha}$. At an exploratory level, a student can vary $\alpha$ to observe corresponding differences in rankings. A more sophisticated exploration might begin with asking whether there are critical $\alpha$ values that change the nature of the ranking in some fundamental way. For instance, $\alpha = 0$ corresponds to strength and $\alpha = 1$ to degree. Also, the range $0 < \alpha < 1$ seems to be fundamentally different from the range $\alpha > 1$. But are there less obvious critical values? See [Opsahl and Skvoretz, 2010] for background on the tuning parameter.

4.2.5. Weights and directions. All of our analysis was conducted on the weighted network under the assumption that a higher volume of emails must have more significance than a lower one. But a simple unweighted graph of email connections, perhaps constructed with some minimum threshold for the number of emails, may reveal information that was obscured by the weighting. Alternatively, students may gain insight from a kind of weighting that treats cc-ed employees differently from our reciprocal square root approach or that assigns importance to emails based on word count or sentiment analysis. And additionally, whether the graph is directed or undirected—that is, whether the sender and receiver are treated symmetrically or not—will result in different outcomes for all the centrality measures, and each may suggest results that the other does not.

4.2.6. A time factor. The majority of the Enron corpus consists of emails from 1998 to 2002. Our graph and corresponding matrices aggregate all the emails into one network. However, it may make sense to consider how the email network changes over time, by month or by quarter. For instance, can an anomaly detection on the network as a time series point out any changes that arose from scandal-related communication? See [Wang et al., 2014] for some work in that direction.

4.2.7. Clustering Extensions. Hierarchical clustering is only one network algorithm that uses adjacencies or distances to break up observations into groups. Partitioning methods typically break the nodes up into groups that partition the units. That is, each node will go into exactly one group. Partitioning Around Medoids (PAM) [Kaufman and Rousseeuw, 1990] iteratively allocates points to the group with the closest medoid (a measure of center based on the nodes themselves), recomputes the medoid, reallocates points, and
repeats until no points need further swapping. Partitioning methods have the disadvantage that the user is required to specify the number of clusters; however, silhouette width can be used to choose the optimal number of clusters [Rousseeuw, 1987].

4.2.8. **Visualizations.** Our research students were particularly interested in different visualizations of the data. They used D3 graphics to create a dependency wheel and an interactive network image (see [http://enron-network.herokuapp.com/TOM](http://enron-network.herokuapp.com/TOM)) [Kaye et al., 2014]. Using applications like Shiny ([http://shiny.rstudio.com/](http://shiny.rstudio.com/)) allows students to think about how best to communicate results, and the Enron data provides myriad opportunities for creative visualizations.

4.3. **Resources.** We have found the following websites useful for further exploration of the data as well as for processed and simplified datasets.

- [https://snap.stanford.edu/data/email-Enron.html](https://snap.stanford.edu/data/email-Enron.html) Stanford Network Analysis Project network analysis and data mining library.
- [http://bailando.sims.berkeley.edu/enron_email.html](http://bailando.sims.berkeley.edu/enron_email.html) UC Berkeley Enron Email Analysis Project, includes natural language processing annotation, visualization and clustering tool, and database representation for efficient querying.
- [http://homes.cs.washington.edu/~jheer//projects/enron/v1/](http://homes.cs.washington.edu/~jheer//projects/enron/v1/) Updated version of visualization and clustering tool by Jeff Heer from Berkeley website above.
- [http://research.cs.queensu.ca/home/skill/otherforms.html](http://research.cs.queensu.ca/home/skill/otherforms.html) Processed forms of Enron data including word frequencies and time stamps.
- [http://cis.jhu.edu/~parky/Enron/](http://cis.jhu.edu/~parky/Enron/) Another set of processed databases into simplified forms like (time, from, to) tuples.

5. **Acknowledgements**

We are grateful to Theo Vassilakis and Jim Addler at Metanautix for their help and suggestions in getting this project started, the Pomona College Math Department for its continued support of undergraduate research, and the students of the Pomona College Undergraduate Research Circle during the Spring semester of 2014, Timothy Kaye, David Khatami, Daniel Metz, and Emily Proulx, for pushing the project to fruition.

6. **Appendix**

As an appendix to this work we provide the dataset given in equation (1). We also provide a list of the 156 employees considered in the analysis (with their departmental affiliation and title). The analysis was done using R ([http://www.r-project.org/](http://www.r-project.org/)) and RStudio ([http://www.rstudio.com/](http://www.rstudio.com/)), and the code used for the analysis is provided as a markdown file and a pdf file.
