QUANTIFYING AND COMPARING CENTRALITY MEASURES FOR NETWORK INDIVIDUALS AS APPLIED TO THE ENRON CORPUS

T. KAYE, D. KHATAMI, D. METZ, E. PROULX

Abstract.
The ever increasing body of social networks creates an opportunity for extensive network analysis and investigations of communications, cliques, and network contributions. In this study, we focus our attention on the Enron email corpus and the corresponding network of employees, attempting to gather information from the email communications. Methods of data reduction on the email corpus were used to create a weighted adjacency matrix in which each i, j-entry corresponds to a weighted count of correspondences from employee i to employee j. While there are many ways to measure importance within a corporate network, of which job title constitutes one such measure, our study focuses on five primary measures: eigenvector centrality, row-sums of a topological overlap matrix, closeness, betweenness, and Opsahl metric. These network analysis metrics were applied to the weighted adjacency matrix to calculate the centrality measures for each individual employee, which were subsequently compiled into ordinally ranked lists of employees for each centrality measure based on decreasing importance. Additionally, the centrality data was visualized using the Data-Driven Documents (D^3) javascript library, allowing for network visualization in terms of department job title and number of emails sent.

In applying the centrality measures to network data, we explore the differences inherent in each measure and work to compare them as well as the corresponding employee importance rankings for each. The metrics in our analysis determined individual importance of employees by applying significant weight to various aspects of the employees' network roles. By identifying employees that are connected to a large number of individuals and simultaneously have extensive correspondences with those individuals, the Opsahl score combines the other measures, proving to be the most useful metric in exploring Enron's inner-corporate structure.

1. Introduction

Social networks are growing at an unprecedented rate. What was once a qualitative undertaking in understanding the sociological aspects of humans has now evolved into a mathematical problem. The growth of the internet has directly impacted our ability to understand social networks [9]. A wealth of information is currently available in a readily accessible digital format, waiting to be analyzed. The data is out there, and the question thus becomes: what can we learn from social network data?

Social network analysis (SNA) is a statistically framed field at heart. We test hypotheses and predictions based on a sample of human interactions. Fields of observations include social media sites such as Facebook, Twitter, and Flickr. Any form of recorded human interaction can be analyzed in some manner. In this paper, we perform quantitative measures on a large email dataset over the span of several years. In such a framework, we are able to view the email dataset using a network theoretic approach, whereby each individual appearing in the dataset becomes a *node* with the network ties corresponding to any type of interaction between individuals. Here we define a *tie* between two nodes as a weighted count of email correspondence. In the case of emails, it is clear that a directed network will result; for a given email, there exists a sender and recipient(s).

Our approach of analyzing the Enron email dataset is through a combination of traditional network theory metrics and visual clustering, which has grown in use in recent times due to its ease-of-use and tactility for the data scientist. Our analysis thus seeks to glean pertinent information from email communications between a subset of people, such as existing cliques between individuals and volume of email activity given a person's role in the dataset.

2. Background

The email dataset we will be analyzing is composed of emails from former Enron Corporation employees[1]. Founded in 1985, the Enron Corporation, an American energy company, quickly became one of the largest natural gas and electricity companies in the world. In November of 2001, Enron filed for bankruptcy after poor accounting and investment practices contributed to hugely significant financial losses. This massive downfall resulted in an extensive legal investigation and a public scandal commonly known as the "Enron Scandal." During the investigation, the Federal Energy Regulatory Commission gained access to a set of emails sent among employees. In May of 2002, this corpus was made publicly accessible online.

In 2004, researchers at SRI International organized and reduced the data set from the initial $619,446$ email messages to $517,431$ messages binned into 150 folders labeled by employee last name and first initial, with emails sent and received between November of 1998 to June of

[1] Code available at `https://github.com/timkaye11/Enron_Network`

2002. This version is devoid of attachments and is referred to as the “March 2, 2004 Version” [12], and is the version that we use throughout our analysis.

3. Research Questions

We ask: who in the network is “important?” We wish to quantify each individual’s importance so that we may compare individuals amongst one another, as well as compare individuals against and across groups. To do so, we must characterize that which aligns with the traditional notion of importance.

In a corporate network, job titles can be considered as indicators of importance. A job title, to some extent, is a measure of one’s role within a company, placing an individual somewhere along a corporate hierarchy. However, job titles are neither easy to quantify, nor do they necessarily act as an accurate gauge of one’s corporate value. To motivate this belief, consider two individuals with the same title. While it’s possible the two are of perfectly equal contributors to the company, it’s rather improbable given that each is likely working on different tasks and interacting with a different set of coworkers. Additionally, if we wish to measure across job titles, we quickly recognize that it is unclear how much more value a CEO offers to a company than, say, a Vice President.

In contrast to job title, we propose to characterize importance by some sense of “connectedness.” An individual who interacts bidirectionally with a significant number of others is, by some measure, a central individual in the context of a network. Someone who is kept in the loop, who connects others, and who generally has high number of correspondences is more likely contributing to the network than an individual who messages only one or two others. This is not to say that the number of correspondences or number of people with whom one connects is directly a measure of importance, but is instead to say that a network is more significantly changed by the removal of a highly connected individual than by the removal a weakly connected individual.

If we are to think of a social network as a graph whose vertices are employees and whose weighted, directed edges correspond somehow with messages sent from some source vertex to some sink vertex, we have a mathematical representation of the network. It is from this representation that one can make measurements and analyses as we attempt to answer how much importance a given individual holds, as gauged through a measure of connectivity, a proxy measure of importance.

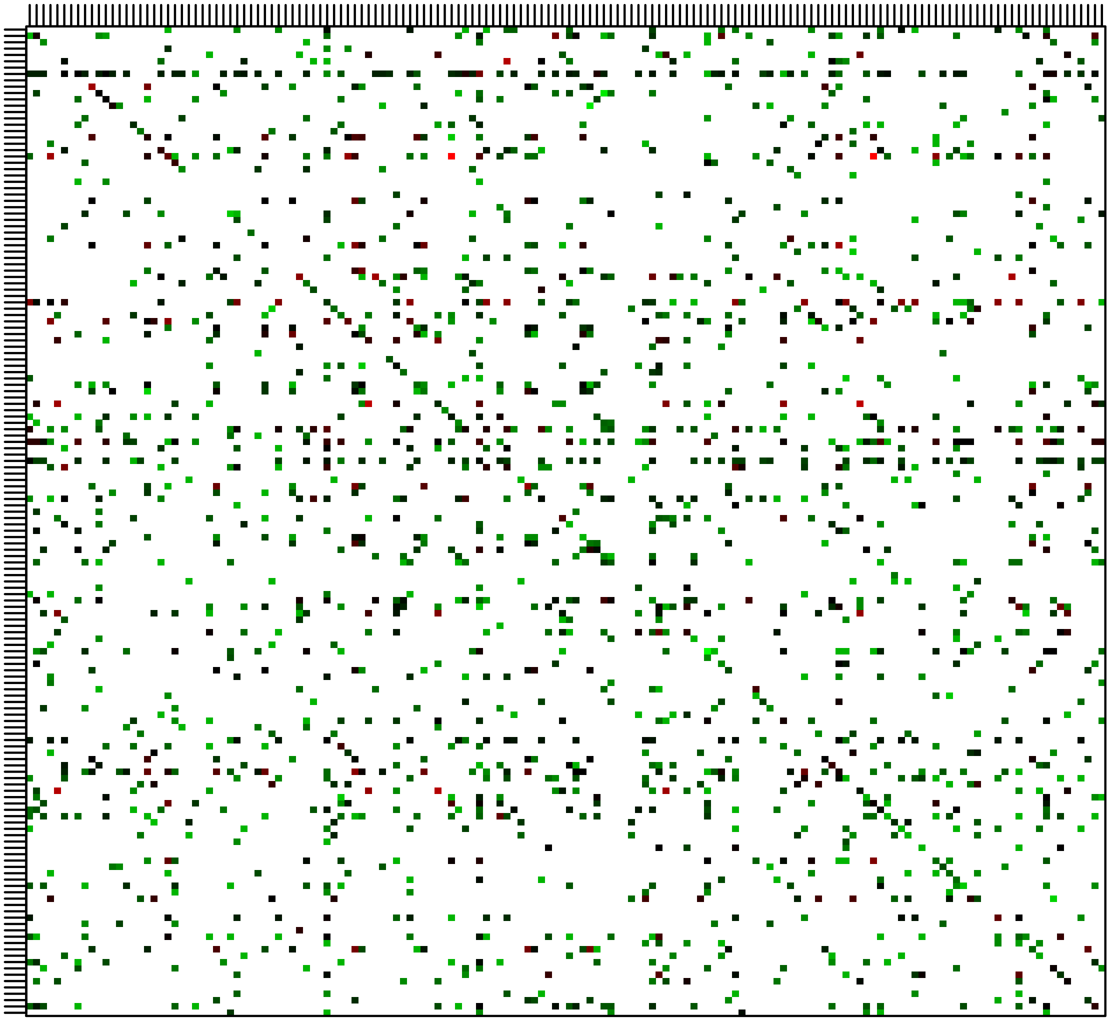


Figure 1. Heatmap of email communication between the 156 individuals, where the column and row correspond to the receiver and sender, respectively. Darker colors indicate more frequent communication, white is no communication. The diagonal represents self-email density. Note that the (i,j) entry differs from the (j,i) entry. The asymmetric nature of the heatmap indicates directionality.

4. Data Reduction

For the given email dataset, we defined our network edges solely as the number of emails between the set of chosen individuals. We therefore ignored the actual body and subject of the email, and narrowed our focus down to the from/to and date attributes of an email. While the body and subject of the email provide useful information in the context of communication, we excluded natural language processing in the scope of this paper. Figure 1 visually represents the frequency of communication between two individuals.

The given dataset appeared as 150 separate email inboxes of the subpoenaed individuals, totalling in size to approximately 18.4 gigabytes of .csv-formatted text. This much data proved too massive to perform efficient data analysis, so the first task in data reduction was optimizing the raw data. R is also incapable of handling more data than the RAM of the computer will allow which, in most cases, is two to four gigabytes. Each email inbox was formatted as one .csv file consisted of 13 columns: *To, ToName, From, FromName, Message, MessageID, Subject, Folder, Filename, Date, Cnt, Cc, Bcc.* Each row of the .csv file corresponded to a single email sent from that inbox. However, it is noted that the initial data was mishandled, with some people with similar names being compressed into a single inbox and some individuals having more than one inbox [12]. This does not prove to be a problem in our analysis, since we are only focused on individual emails and not how the email datasets were binned. Thus, the data reduction task came down to generating a large matrix of emails' *From, To, Date,* and *Cc* attributes.

We chose the R package `ff` for our large dataset handling [1], which provides memory-efficient storage and fast access to large data. Using the supporting package `ffbase`, we converted each .csv file into an `ff` data frame, which optimizes data analysis by efficiently indexing and organizing the .csv information. In doing so, what was originally 18.4 gigabytes of data was compressed into 329 *megabytes*, a factor of 56 smaller in terms of disk usage and compact enough to be handled in R. The package `ff` accomplishes such a drastic optimization of large data by mapping only a section of the given data into the main memory and providing efficient data querying of a given data frame. The file compression works through exploiting sparse file allocation techniques in physical memory. As mentioned earlier, our attributes of focus were only those directly linked to from/to and date information of an email so the data corresponding to the subject information and actual email content were trimmed.

The end-goal of the data reduction pipeline is to produce a weighted adjacency matrix on which we can apply our social network analysis metrics. We then created a list of 156 aliases for individuals in the email dataset and constructed a 156 × 156 square adjacency matrix, whereby the row and column corresponds to the sender and recipient of a given email, respectively. Each i, j-entry in the matrix corresponds to a weighted number of times individual i communicated with individual j. The weighting scheme is as follows: first we iterate through each email in the corpus. Suppose individual i sends an email to person j, with cc-ed individuals $C = \{c_1, c_2, \ldots, c_n\}$. Then we add a count of one to the i, j-th entry of adjacency matrix A:

$$A_{ij} := 1 + A_{ij}.$$

Yet we also wish to count the communications between sender i and the individuals in the Cc. However, we wish to quantify the importance of a communication link, so we created a weighting scheme, whereby a certain communication is weighted as *less important* if it has more individuals appearing in the Cc. The weighting scheme thus becomes, for sender i communicating with cc-ed individual k,

$$A_{ik} := (1 + n_{cc})^{-1/2} + A_{ik},$$

where n_{cc} is the number of individuals appearing in the Cc list. Thus, a few Cc's will still keep an individual with some importance, but the importance of a communication link will decrease at an inverse square-root rate.

We implemented a searching algorithm using a table from [12] that listed all known aliases of the 156 individuals of interest, as well as some of their departments, which will be discussed later. The search looked at the to/from/Cc fields as a string and, using the *grep* function in R, if one of the alias emails appeared as a subfield, recorded a find for that individual.

The final weighted-directed adjacency matrix that we generated revealed several of the 156 individuals who, in the context of the communication network consisting of the other 156 individuals, found to be insignificant/irrelevant in the constructed social network i.e. they sent little to no emails to any of the other individuals. Thus, a total of 16 individuals were found to have little to no significance. A heatmap representation of the weighted adjacency metric can be found in Figure 1. Notice how certain rows are denser than others, indicating that the corresponding individual emails frequently.

5. Methodology

Aliases: For the email addresses contained in the Enron corpus, we focus solely on the addresses with the Enron domain. While there were some outliers, we found that most people share the same permutations of their first, middle and last name. Six of the generalized formats are:

$$< \text{first initial} >< \text{last name} > \text{@enron.com}$$
$$< \text{first initial} >< \text{last name} > @ < \text{dept} > \text{enron.com}$$
$$< \text{first name} > . < \text{middle initial} > . < \text{last name} > \text{@enron.com}$$
$$< \text{first name} >< \text{last name} > \text{@enron.com}$$
$$< \text{middle initial} > .. < \text{last name} > \text{@enron.com}$$
$$< \text{last name} > - < \text{first initial} > \text{@enron.com}$$

5.1. Measures of Centrality.

For the Enron employees in our analysis, we use the following network analysis metrics to measure individual employee centralities and rank the employees based on importance for each measure.

5.1.1. *Eigenvector Centrality.*

Eigenvector Centrality assigns a value to each vertex of a graph defined by an adjacency matrix. It is motivated by the belief that one's degree, rather than measure simply one's number of connections, should instead be a weighted measure of one's connections in that a connection from a "well-connected" individual is worth more than a connection from a "poorly-connected" individual. Alternatively, from *Social Network Analysis for Startups* [10]:

> " Eigenvector centrality is like a recursive version of degree centrality. The algorithm works roughly as follows:
>
> 1. Start by assigning a centrality score of 1 to all nodes ($v_i = 1$ for all i in the network).
> 2. Recompute the scores of each node as a weighted sum of centralities of all nodes in a node's neighborhood:
>
> $$v_i = \sum_{j \in N} x_{i,j} \cdot v_j$$
>
> 3. Normalize v by dividing each value by the largest value.
> 4. Repeat steps 2 and 3 until the values of v stop changing. "

So we see that eigenvector centrality is a measure of "connectedness," giving greater values to those who are connected to more people, and weighting the value of those connections based the "connectedness" of the connector. Because little value is given to an incoming connection from an isolated individual and great(er) value is given to an incoming connection from a "central" individual, any individual's importance is dependent upon both the number of connections he or she has *and* those connections are weighted in accordance with the importance of the senders. Those most highly ranked will be those with a large number of connections and whose connections themselves are connected with a large number of others. As an added benefit, calculating the central eigenvector (as through the `evcent` function of the `sna` [4] package for R) accounts for directionality, which means one can also compare how important someone is as a sender in contrast to how important someone is as a receiver.

5.1.2. *Topological Overlap Matrix.*

A topological overlap matrix, often abbreviated TOM, is a matrix whose i, j entry is a measure of the connectedness of network vertices i and j. To calcualte the entries of a TOM, we must first define a neighborhood; A neighborhood of a vertex k is the set of vertices adjacent to vertex k. In a TOM, the value of the i, j entry corresponds to the size of the intersection of the neighborhoods of i and j, divided by the size of the smaller neighborhood (of either i or j). In the context of our social network, TOM will give greater weighting to individuals whose social circles significantly overlap.

To calculate a TOM matrix [5], let us begin with a matrix $A = [a_{i,j}]$ where $a_{i,j}$ is equal to the number of emails sent between person i and person j (summed across both directions), divided by the largest number of emails exchanged between any two individuals. This means that A is both symmetric and has all entries between 0 and 1.

Given $A = [a_{i,j}]$, we calculate our TOM adjacency matrix as

$$\text{TOM} = [\text{TOM}_{i,j}] = \frac{\sum_{u \neq i,j} a_{i,u} a_{j,u} + a_{i,j}}{\min\left(\sum_{u \neq i,j} a_{i,j}, \sum_{u \neq i,j} a_{j,u}\right) + 1}$$

From this TOM adjacency matrix, we wish to assign a single value to each individual. We do so by taking the row-sum for each individual. The resulting vector, TOMrank may therefore be calculated as

$$\text{TOMrank} = [\text{TOMrank}_i] = \sum_j \text{TOM}_{i,j}$$

5.1.3. *Closeness.*
Looking into the closeness of nodes within the network is important in identifying the nodes that can reach others quickly. While closeness is rendered moot for graphs with various connected components, we see from Figure 2 that the communication seems to be centered around one central cluster.

Social networks are often conveniently denoted as $G = (V, E)$, where the graph G is composed of vertices V and edges E. In the context of the Enron corpus, each of the 156 employees is a vertex, while the adjacency matrix defines the edges between the nodes. We assume that for each established edge, the weight $w > 0$.

Considering $i, j \in V$, we can define $d_G(i, j)$ to be the length of the shortest path between a and b. This distance is calculated using *Dijkstra's Algorithm* [3]. Additionally, $\sigma_G(s, t)$ is the number of shortest paths between a and b. For example, if i sends an email to itself, then $d_G(i, i) = 0$ while $\sigma_G(i, i) = 1$.

We can define closeness for a given vertex i in a directed (non-symmetric) network as:

$$C_G(i) = \left(\sum_{j \in V} d_G(i, j)\right)^{-1}$$

5.1.4. *Betweenness.*
Betweenness is another centrality method within networks. While betweenness is a measure solely for undirected networks, it is a fundamental concept for the analysis of social networks [11]. Betweenness looks at how central a node is with respect to the communication of all other nodes. We incorporate Brandes' [3] algorithm which takes weighted edges into consideration. This approach takes into account that the communication between two nodes may be quicker along more heavily weighted edges as opposed to fewer, untraveled edges.

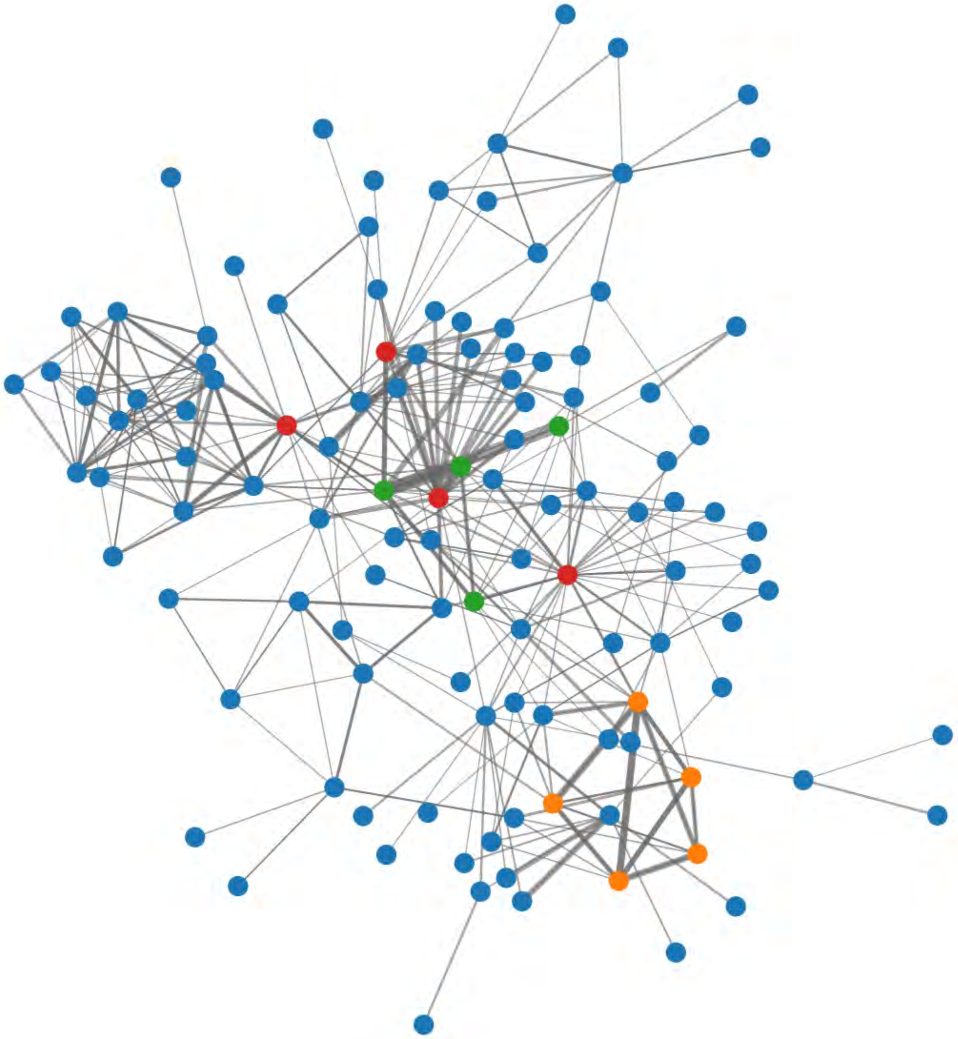


Figure 2. Visual representation of the Enron network. The non-blue colors represent the top individuals for each metric. Red shows betweenness, green TOM/closeness, and evcent orange. The visualization reveals three major cliques.

The Brandes algorithm also has proven to be more computationally efficient, as it reduces the complexity from $\mathcal{O}(mn + n^2 \log n)$ to $\mathcal{O}(mn)$, where m is the total number of edges, and n is the number of rows of the adjacency matrix. However, Opsahl found that Brandes [3] generalized algorithm focuses on the total number of emails sent, and fails to consider the connections between nodes" [8]. Thus, we incorporated Opsahl's betweenness algorithm, found in the R package `t-net`[7]. To transform the adjacency matrix to be undirected, we simply added the transpose and set the diagonal equal to zero.

The betweenness for a vertex v_i is denoted as:

$$B_G(i) = \sum_{j \neq i \neq k \in V} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

where $\sigma_{jk}(i)$ refers to the number of shortest paths between v_j and v_k in which v_i lies on the path, and σ_{jk} is the number of shortest paths between v_j and v_k.

5.2. **Opsahl Measure.**

The degree and strength measures of a network are amongst the most straightforward. The degree is calculated by finding the number of people that a given employee sent emails to (the row sum of the binary adjacency matrix). Moreover, the strength is simply the row sum of the weighted matrix, or the total number of emails sent by an employee. While these metrics seem to just scratch the surface, Opsahl [8] posed a new measure that combines degree and strength of a network. This metric incorporates a tuning parameter α, which establishes the relative importance of the number of ties compared to the tie weights (Opsahl 2010) and can be generalized as follows :

$$C_O(i) = \text{degree}^{1-\alpha}\text{strength}^{\alpha}$$

where $C_O(i)$ is the Opsahl measure with the α parameter for any node i. In our application of Opsahl's measure, we used an alpha parameter of 0.5 as it evenly weights degree and strength.

5.3. **Theoretic Comparison of Centrality Measures.**

Each measure of centrality assigns weighting differently.

- Eigenvector centrality values a vertex v_i recursively based upon the weights of incoming edges from v_j to v_i and upon the centrality values of those source vertices $v_j \in V$. Because of this, eigenvector centrality will assign high values to individuals who are connected with a large number of individuals who are in turn connected with a large number of individuals.
- A topological overlap matrix (TOM) is a symmetric matrix whose i, j entry is loosely a measurement of the size of the intersections of the neighborhoods of v_i and v_j divided by the size of the smaller neighborhood. The row-sum of a TOM matrix is then a measure of how well connected an individual as compared to others. Those with a large number of connections will rank most highly.
- The closeness value for a given vertex v_i is roughly a measure of its independence and efficiency. As closeness cannot be calculated for networks with unconnected components, it leads to very small values for vertices that send few emails to a select

number of people. Thus as a result, vertices who rank high on the closeness measure are able to efficiently communicate with other vertices.

- Betweenness centrality measures the extent to which a vertex v_i lies between other vertices on their geodesics (shortest paths). Vertices that rank high in betweenness centrality are capable of influencing the spread of information through the network, by facilitating or altering the communication between other vertices. However unlike closeness, betweenness is not a local measure of centrality, and those who rank high in betweenness therefore have the potential to create influence both directly and indirectly.
- The Opsahl metric combines both closeness and betweenness, by weighting each one more strongly depending on a tuning parameter. With a proper tuning parameter, the metric will rank individuals who are both central and well-connected to other individuals in the network.

Note that among the metrics we consider, only closeness and degree centrality focus consideration upon direct connections. Therefore, we expect that these measures have slightly higher correlation than the other measures.

6. Results

In order to visualize the Enron Email network, we used D^3 (Data-Driven Documents) - a Javascript library, to create interactive visualizations of the data (D^3 visualizations frequently appear in the *New York Times*). As it is difficult to gain an understanding of the network based just on the matrix, the visualizations allow one to view the network in terms of department, job title and number of emails sent. Figure 3 (the dependency wheel) displays the two-way communication of the Enron employees. In this figure, the names of the employees appear along the outer radius, where the color of the segment indicates the department. The thickness of each person's section indicates the number of emails sent, in comparison to the other employees. The chords within the diagram indicate emails sent/received between two given employees. For example, an initial look at Figure 4 indicates that *Jeff Dasovich* is the largest sub-section of the graph, meaning that he sent the most emails amongst the employees. In addition, the communication within one's own department was removed, so that we could highlight the inter-department communications within the directed network. An interactive version of the graph as available at `http://obscure-meadow-3612.herokuapp.com`.

Moreover, Figure 2 presents the data as a force-directed graph, using the D^3 javascript library. This figure considers all the connections in the network, and has the added benefit of being able to determine how central a vertex appears in the force-directed layout. The *force-directed* layout of D^3 uses Verlet Integration and Barnes-Hut approximation, as well as a repulsive charge force to keep vertices centered in the visible area [2]. The legend

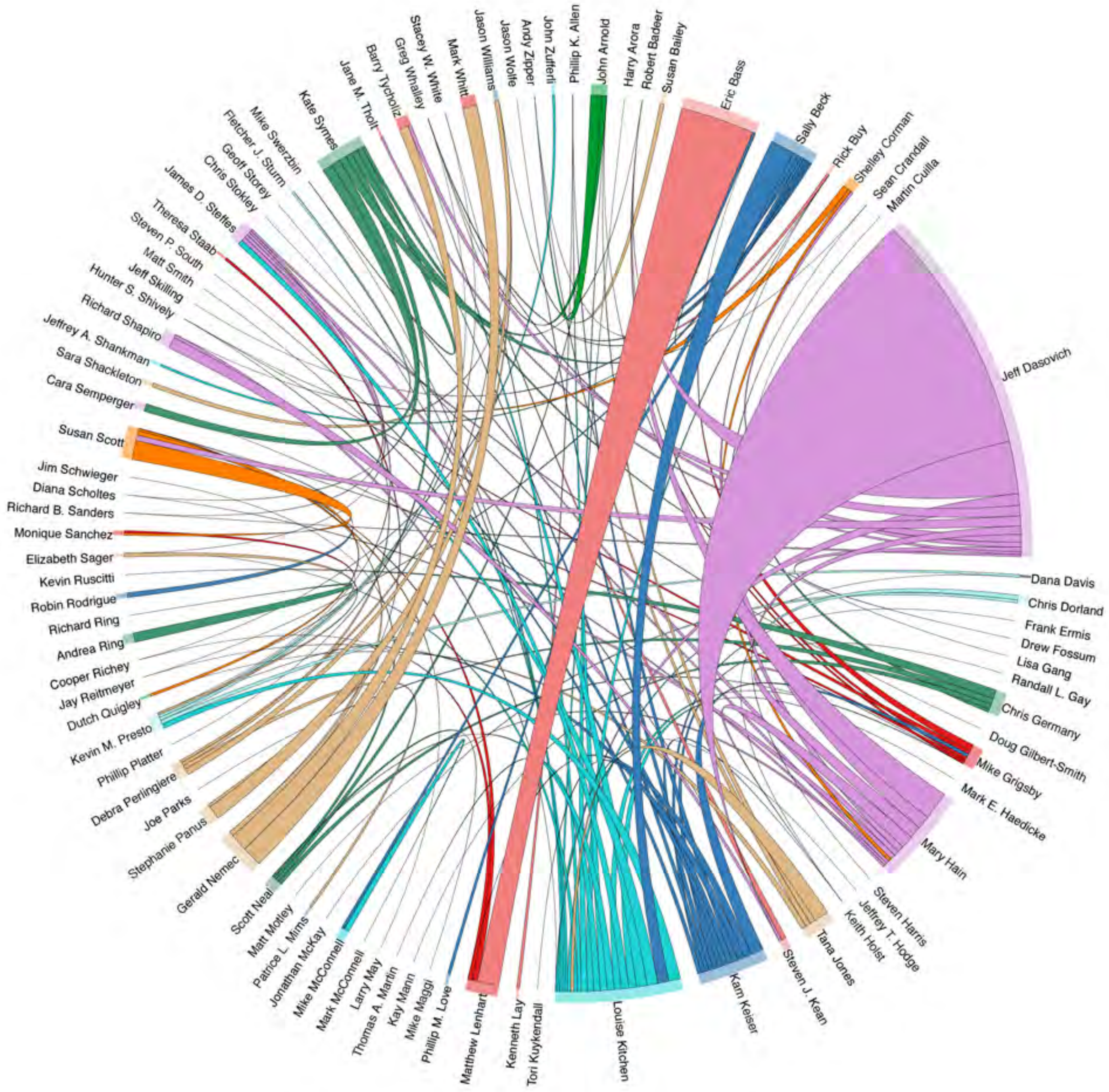
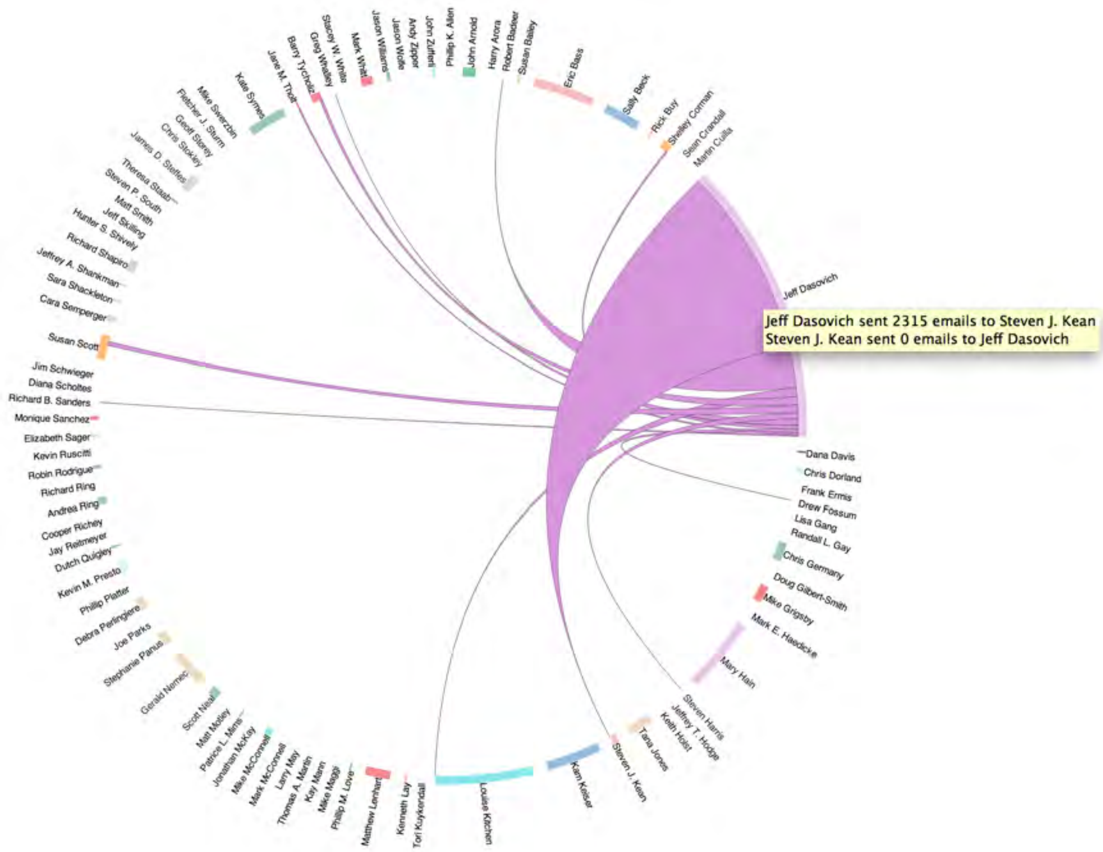


FIGURE 3. Dependency wheel, showing all the ties of the 156 individuals. Each tie is weighted on both ends with respect to how much each person sends to the other person.

indicates the top 10 vertices in the graph, according to their distance with respect to the origin. A complete/interactive version of this figure can be seen at `http://enron-network.herokuapp.com/TOM`.

Note that for the construction of these figures, we found that the inclusion of all 156 original vertices in the adjacency matrix lead to a cluttered representation. Thus, we removed all the rows/columns with only zeros, and set a threshold of 25 emails sent/received. This



Jeff Dasovich sent 2315 emails to Steven J. Kean
Steven J. Kean sent 0 emails to Jeff Dasovich

FIGURE 4. Highlight of the dependency wheel of Jeff Dasovich, showing only his connections and relative strength with respect to the others.

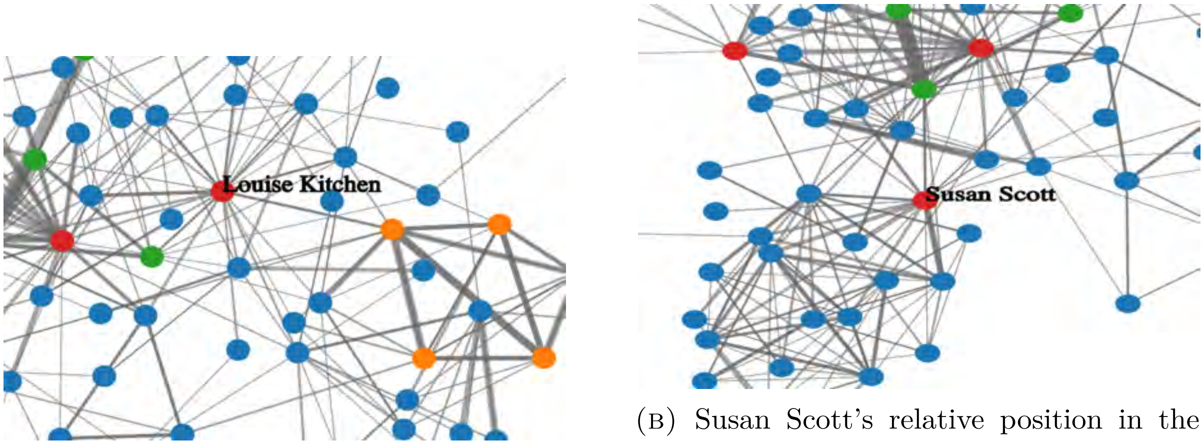
filter limited the network to 84 subsections for the Dependency graph, and 114 vertices for the force-directed graph. This allowed for a more elegant layout for Figures 3 and 2.

We also have a very unusual case where not only do two individuals, Mark Dana Davis and Dana Davis, share the same email account folder, but for some time they both used the email address dana.davis@enron.com. To resolve this issue, we simply combined the two addresses, as both employees had rather small email records relative to the entire corpus. [6].

6.1. **Top Ranked Individuals for each Metric.**

Based on each of our centrality methods, we determined the 10 highly ranked employees. For the purposes of comparison, we then found the corresponding department for each of these employees. Upon inspection, we see that no two methods led to the same results, yet 'Jeff Dasovich' appears at the top of the list in three of the five Centrality methods. This result coincides with Jeff Dasovich's magnitude within the D^3 visualization.

Upon reviewing the D^3 graphs in conjunction with the results from the centrality measures, we were able to decipher key relations and observations in the network. Notably, the thickest



(A) Louise Kitchen, at the heart of the cluster.

(B) Susan Scott's relative position in the graph, showing a high betweenness between cliques (a subset of individuals who communicate regularly amongst themselves).

FIGURE 5. A zoom-in of the graph, showing the persons of interest and their relative position.

line in Figure 4 is between Jeff Dasovich - the Director of Regulatory and Government Affairs, and Steven Kean - the Chief of Staff. A closer look into the communication between Dasovich and Kean reveals that of the 3418 non-inter-department emails the Dasovich sent, 2315 of them were sent to Kean. On the other hand, of the 73 emails that Kean sent, none of them were directed to Dasovich. This suggests that Dasovich and Kean shared a one-sided relationship, indicating that the emails were more informative than personal. In addition we see from Table 1 (the centrality table) that Kean ranked amongst the top five for the *TOM* and *closeness* measures. We also note that Kean does not appear in the top twenty for the other three measures. While Kean sent very few emails, his high rank in the *TOM* and *closeness* measures can be justified by the strong one-sided connection with Dasovich, as *TOM* runs on a symmetric matrix. Furthermore, from looking at Figure 4, we see that Kean's central position within the graph is primarily attributed to Dasovich's centrality, as Kean sent/received emails from five people, with the majority of those being from Dasovich.

From Figure 5(A) we observe that Louise Kitchen appears close to the center of the network, as the D^3 force-directed layout places central vertices near the center of the page. In addition, along with Dasovich, Kitchen was the only employee who ranked in the top-ten for all five categories. However unlike Dasovich, the majority of Kitchen's emails appear to be two-way communications - meaning that regardless of symmetry, Louise Kitchen is shown to be a central employee. This result coincides with the fact that Kitchen operated as the COO of Enron.

From the betweenness metric, we see the Susan Scott ranks in the top five, yet she fails to rank in the top five for the other metrics. Though she did not send a large amount of emails, she links a cluster of the graph with the primary component - acting as a conduit. Figure

5(B) indicates that though Scott communicated with various departments, she was also the key liaison between Dasovich and the sub-cluster within the graph.

From the metrics used, we found a series of “cliques.” Our analysis of the `evcent` metric reveals a centrality focused on relative tie strength between a connected subset of individuals, or a “clique,” regardless of their relevance to the whole cluster (Figure 2). The top five individuals using the `evcent` metric are, in order, Tana Jones, Sara Shackleton, Stephanie Panus, Marie Heard, and Susan Bailey. It is at first glance surprising that the top five individuals all work for the legal department (Table 2), but fits within the framework that `evcent` reveals a *single* clique. In this case, we find that legal department clique is ranked very highly, in which the top five members communicate inside of the department with a very high email traffic. As shown earlier, we consider TOM/closeness in the same vein due to their strong correlation (Figure 6). Our measure with the TOM/closeness metric reveals a clique that is most central to the whole cluster, the Regulatory and Government Affairs department of Enron (Table 2). Intuitively, a department that focuses on inter-corporate and governmental communication linkings would appear highly in our TOM/closeness measure. The most obvious metric to interpret is betweenness which, unsurprisingly, ranks those individuals who bridge cliques highly. Thus, we find that individuals in a betweenness metric are ranked by how much they pull cliques into the center. Finally, we analyze the Opsahl metric measure on the cluster with an Opsahl parameter of $\alpha = 0.5$. Opsahl helps find a nice balance between interconnectedness and relative communication channel strength, i.e. how efficient an individual is at acting like a traffic of communication. Thus, those individuals ranked highly in the Opsahl metric act as a bridge to many individuals and appear central to the cluster. In our interpretation of importance in the network we constructed, the Opsahl metric proves to be the most useful measure in examining the inner-corporate structure of Enron by considering not just centrality, but the cohesiveness of the cluster due to the individual.

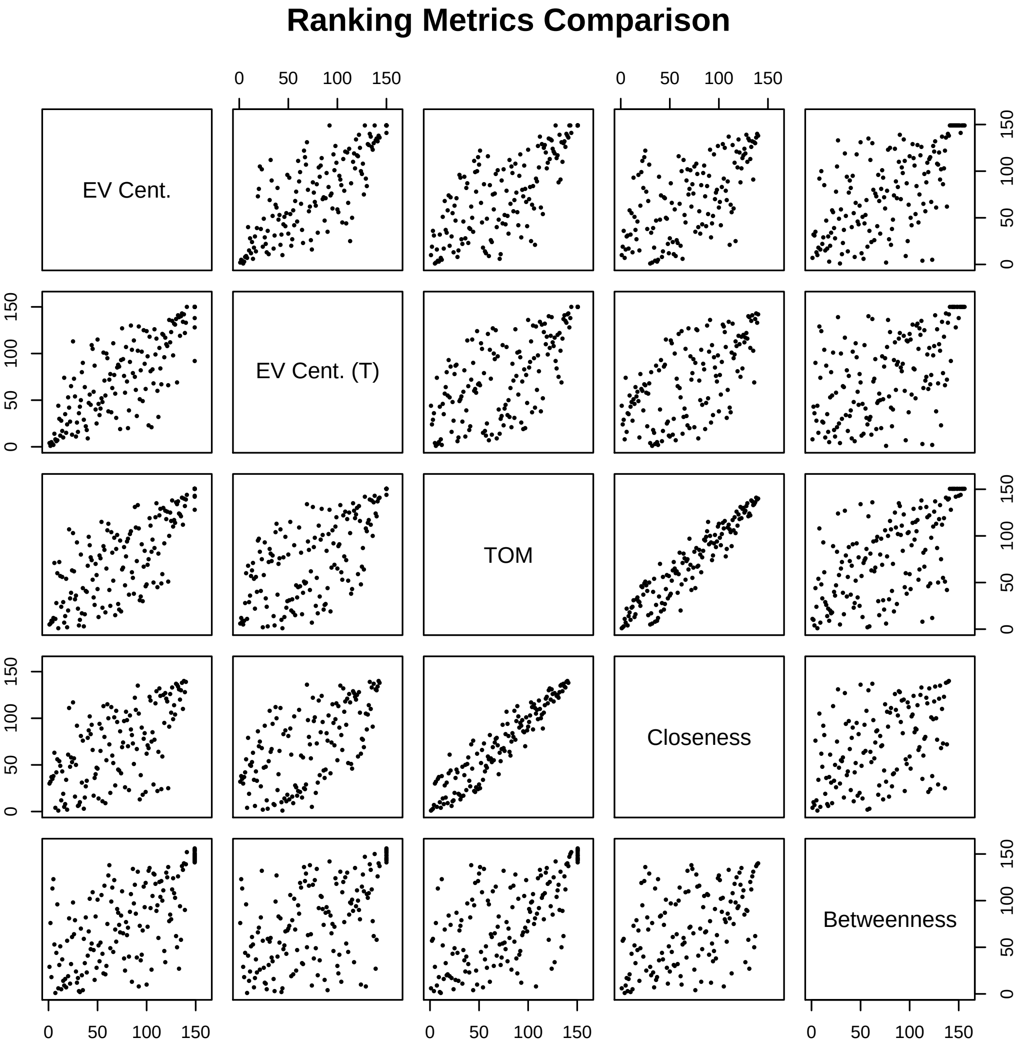


Figure 6. Pairs plot comparison of how the individuals are ranked in a given metric. EV Cent. is the eigenvector centrality measure; EV Cent. (T) is the transpose; TOM is the topological overlap measure.

7. Discussion

While the Enron email corpus provides an opportunity for network analysis on real data, the incomplete and complex nature of the public email messages poses some limitations to our research. Although Enron employed thousands of staff members before its bankruptcy, the unaltered version of the corpus only consists of 150 email folders and 156 unique employee mailboxes, suggesting that the data may not be a complete representation of the Enron network. In addition, the reduced dataset we used was missing some messages that were deleted from the corpus as requested by Enron employees [12], and it is unclear how many email messages were flushed from mailboxes before they were subpoenaed.

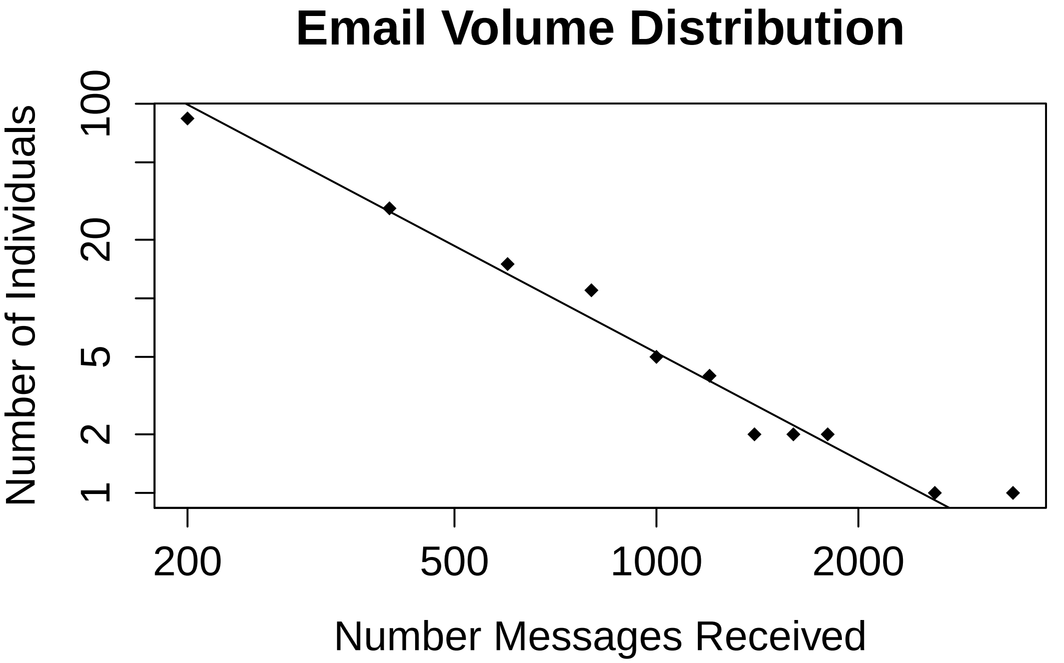


Figure 7. Distribution of the number of emails an individual receives. The distribution of how many individuals receive a certain amount of emails exhibits a power-law like distribution. A linear regression yielded an R-squared value 0.978 and a t-value of -17.78 for the coefficient.

As mentioned, in the process of reducing and organizing the emails, we searched for aliases and filtered emails based on six general address prefix formats and the address suffix `@enron.com`, with the exception being the address suffix `@<dept>enron.com`. However, many employees used multiple email addresses with a different suffixes, for example Phillip K. Allen utilized `pallen@enron.com`, as well as `pallen70@hotmail.com` and `pallen@hotmail.com`. Our filtering method focused only on those aliases containing `@enron.com`, resulting in the exclusion of some addresses and corresponding messages. As our measures of centrality and importance rely on the number of emails sent, received, and cc-ed for each employee, the number of messages and email aliases included alters our centrality calculations.

When cleaning up the data, we initially attempted to show changes in employees' centralities over time using month-binned matrices of email counts. With such sparse matrices, the centrality measures being used were not providing enough information from which we could glean a conclusion. We ultimately decided to focus our attention instead on various measures of connectivity as they relate to each other, compiling the email counts across the entire relevant time period into a single matrix to get more accurate measures of centrality. A next step may be to segment message counts by time in increments of three months and track changes of importance of people from various Enron departments throughout the scandal. Our centrality measures provide flags for individuals of interest that would then be followed up with qualitative analyses and the cause of their importance, such as an examination of job title and its role in the network.

Although we emphasized individual centrality in the network, the Enron email corpus not only acts as a lens through which to investigate network dynamics, but can also be used in future research to explore message content, sentiment analysis, predicting email responses, and several other message content-related analyses. Figure 7 shows a power-law scaling of the number of individuals who receive a certain amount of emails, as is expected for a scale-free network such as the Enron email corpus. These power-law networks are common in social networks, so the methodology adopted in this paper can readily be applied to related networks such as Twitter, LinkedIn, and many others.

While we selected five different centrality algorithms, there may have been additional room for alternative algorithms. For instance, consider the network as a whole and some measure of whole-group connectivity. If we can remove one individual from the group and recalculate this connectivity measure, we can assign this individual some value indicating their contribution to the group's connectivity as a whole. By repeating this process for each individual in the group, measuring the impact of one's removal from a group in terms of the group's remaining level of connectivity, we are left with another measure of centrality. This sort of approach would be more applicable in an email/company network, as opposed to a social network. We can distinguish an email network such as the Enron Corpus from a social network by considering the prior knowledge of the dataset. For instance, while we could look up an employee's position and department to gauge their relative importance in the Enron Corpus, such information would be less useful in a social network.

8. Conclusion

In an investigation of the Enron email corpus, we sought to answer the question of "who in this social network is most important?" Through our analysis, we noted that looking beyond job title as a gauge for determining network importance, there exist a number of measures that can be used to quantify the "connectivity" of individuals in a social network. Each metric attributed significant weight to different aspects of one's network role in order to determine individual importance. On their own, none of these metrics captured our intuitive understanding of social network importance. A socially important individual is one who is both connected to a large number of individuals *and* who has significant correspondences with those individuals. The Opsahl measure does just that, offering a weighted combination of the closeness and betweenness measures. Individuals ranked as highly important by the Opsahl metric are both central to the cluster and serve as a bridge to many individuals.

In addition, the ability of the metrics to identify cliques could be applied to other social networks. Unlike the corpus where we had information on the departments of individuals, such metric analysis could be used to identify common characteristics between individuals.

9. Acknowledgments

We would like to thank Ghassan Sarkis, Associate Professor of Mathematics at Pomona College, and Jo Hardin, Associate Professor of Mathematics at Pomona College, for their guidance and patience in navigating us through a multifaceted data scientific problem. We would also like to acknowledge the Pomona College Mathematics Research Circle Program for continual support in undergraduate mathematics research.

10. Appendix

Table 1. Top 10 Individuals for Each Metric

Evcent	TOM	Closeness	Betweenness	Opsahl
Tana Jones	Jeff Dasovich	Jeff Dasovich	Louise Kitchen	Jeff Dasovich
Sara Shackleton	Richard Shapiro	Richard Shapiro	Jeff Dasovich	Mike Grigsby
Stephanie Panus	Steven J. Kean	Steven J. Kean	Mike Grigsby	Sally Beck
Marie Heard	Mary Hain	Louise Kitchen	Susan Scott	Louise Kitchen
Susan Bailey	James D. Steffes	Mary Hain	Mary Hain	Tana Jones
Kay Mann	Robert Badeer	Barry Tycholiz	Scott Neal	Susan Scott
Louise Kitchen	Louise Kitchen	Mike Grigsby	Robert Badeer	Chris Germany
Elizabeth Sager	Susan Scott	James D. Steffes	Kate Symes	Sara Shackleton
Jason Williams	Barry Tycholiz	Robert Badeer	Sally Beck	Kam Keiser
Jeff Dasovich	Greg Whalley	Susan Scott	Monique Sanchez	Mary Hain

Table 2. Top 10 Departments for Each Metric

Evcent	TOM	Closeness	Betweenness	Opsahl
ENA Legal	RGA	RGA	EWS	RGA
ENA Legal	RGA	RGA	RGA	ENA Gas West
ENA Legal	Enron	Enron	ENA Gas West	Energy Operations
ENA Legal	RGA	EWS	ETS	EWS
ENA Legal	RGA	RGA	RGA	ENA Legal
ENA Legal	ENA West Power	ENA Gas West	ENA Gas East	ETS
EWS	EWS	ENA Gas West	ENA West Power	ENA Gas East
ENA Legal	ETS	RGA	ENA West Power	ENA Legal
ENA Gas Central	ENA Gas West	ENA West Power	Energy Operations	Energy Operations
RGA	EWS	ETS	ENA Gas West	RGA

References

[1] D. Adler, C. Gläser, O. Nenadic, J. Oehlschlägel, and W. Zucchini. *ff: memory-efficient storage of large data on disk and fast access functions*, 2014. R package version 2.2-13.

[2] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309, 2011.

[3] U. Brandes. A faster algorithm for betweenness centrality*. *The Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[4] C. T. Butts. *sna: Tools for Social Network Analysis*, 2014. R package version 2.3-2.

[5] A. Li and S. Horvath. Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics*, 23(2):222–231, 2007.

[6] G. Namata, L. Getoor, and C. Diehl. Inferring formal titles in organizational email archives. In *Proc. of the ICML Workshop on Statistical Network Analysis*, 2006.

[7] T. Opsahl. *Structure and Evolution of Weighted Networks*. University of London (Queen Mary College), London, UK, 2009.

[8] T. Opsahl, F. Agneessens, and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, 2010.

[9] E. Otte and R. Rousseau. Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science*, 28(6):441–453, 2002.

[10] M. Tsvetovat and A. Kouznetsov. *Social Network Analysis for Startups: Finding connections on the social web*. " O'Reilly Media, Inc.", 2011.

[11] D. R. White and S. P. Borgatti. Betweenness centrality measures for directed graphs. *Social Networks*, 16(4):335–346, 1994.

[12] Y. Zhou. Mining organizational emails for social networks with application to enron corpus, 2008. Copyright - Copyright ProQuest, UMI Dissertations Publishing 2008; Last updated - 2014-01-21; First page - n/a; M3: Ph.D.

Department of Mathematics, Pomona College

E-mail address: timothy.kaye@pomona.edu, david.khatami@pomona.edu, daniel.metz@pomona.edu, emily.proulx@pomona.edu, jo.hardin@pomona.edu, ghassan.sarkis@pomona.edu