Data Mining Academic Emails to Model Employee Behaviors and Analyze Organizational Structure

Kayla M. Straub

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Electrical Engineering

Robert W. McGwier, Chair

Aloysius A. Beex

Michael M. Buehrer

Bert Huang

March 25, 2016

Blacksburg, Virginia

Keywords:

Copyright 2016, Kayla M. Straub

Data Mining Academic Emails to Model Employee Behaviors and Analyze Organizational Structure

Kayla M. Straub

(ABSTRACT)

[This is just the paper abstract - I plan to rewrite this near the end]. Email correspondence has become the predominant method of communication for businesses. If not for the inherent privacy concerns, this electronically searchable data could be used to better understand how employees interact. For example, after the Enron dataset was made available, researchers were able to provide great insight into employee behaviors based on the available data despite the many challenges with that dataset. The work in this paper demonstrates the application of a suite of methods to an appropriately anonymized email dataset created from volunteers' email metadata. This new dataset, from an internal email server, is first used to validate machine learning and feature extraction algorithms and then to generate insight into the interactions within the center. Based solely on email metadata, a random forest approach modeled behavior patterns and accurately classified employees by job title. The algorithm performed very well not only on participants in the study but also on other members of the center who were connected to participants through email. Furthermore, the data revealed relationships not present in the formal operating structure. The result is a much fuller understanding of the center's internal structure than can be found in the official organization chart.

Dedication

Acknowledgments

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

A reorganization of a business can be very costly and has great effects once implemented, either for better or for worse. While this official hierarchy is important, there is an equally important organic organization of any business, which may or may not be reflected in the official organization chart. Understanding this unofficial structure is important, but due to its informal nature, it can be difficult to determine. One massive source of electronically searchable information that could be used to better understand this hidden structure is the business's emails. However, privacy concerns inhibit most research thrusts into email analysis.

If access was given to email metadata, there could be several potential applications to this information beyond gaining a better understanding of the organizational structure. It could be possible

ot identify people of influence or potential leaders. By recognizing changing behavior, potential insider treats could be identified.

1.2 Problem Overview

The study presented in this paper collected an appropriately anonymized email metadata dataset to demonstrate to what extent this metadata can be used to determine an organic organizational chart. Furthermore, this research measured how well this organic organizational chart matches the official organizational chart. The study utilized email metadata from 37 voluntary participants. This metadata is used not only to analyze the voluntary participants, but also to what extent the non-participant members of the organization can be characterized.

1.3 Approach

The 98 features used in this study can be grouped into two common areas in email analytics: traffic-based and social-based. Using random forests, these features are used to predict each employee's job title. The ability to accurately determine job title using as few as three features is demonstrated. Finally, a comparison between relationships displayed in the data with the formal organizational chart is presented.

1.4 Contributions

The contributions outlined in this thesis are broad. First, a new, fully anonymized dataset was developed from raw emails from an academic research environment. This dataset, unlike others before it, has accurate job title labels. Furthermore, a unique combination of features were calculated from this data. Many had been used in email analysis before, but some are new and specific to this data. In general, the problem of job title classification has not been extensively studied. However, the results in this thesis surpass the results of previous research.

1.5 Outline

This paper continues by discussing the related works in Section 2. Section 3 describes the process of data collection and some statistics of the dataset and the features extracted from the data. The methods investigated using these features are covered in Section 4. The results of the analysis are presented in Section 5. Section 6 presents opportunities for future work, and Section 7 concludes the thesis.

Chapter 2

Prior Work

2.1 Background

Email is a pervasive medium for communication in modern society—particularly in the workplace. In 2015, there were over 2.6 billion email users [1]. It is projected that by the end of 2019, over one third of the global population will be using email. In fact, the average business email user sends and receives a total of 112 emails per day. Corporate email alone accounts for 54.7% of worldwide email traffic. Retention of large email archives has become common practice with decreasing memory size and cost [2]. Out of 600 employees at a high-tech company, the average employee had 28,660 emails stored in 133 folders which is a significant increase over the past ten years. This is a considerable amount of untapped information that could be leveraged to characterize employee roles within an organization.

2.2 Commonly Used Datasets

Since the Enron email dataset was released in 2004 [3], it has been extensively researched on topics including spam classification [4], [5]; email categorization [6], [7]; and recipient prediction [8], [9]. However, there are known flaws and discrepancies with even the most recent versions of this dataset—ranging from misspelled email addresses [10] to duplicate email addresses [11], and misfiled emails [12]. In one of the most popular forms of the dataset [13], the database includes 253,735 emails sent as "CC" and 253,713 emails sent as "BCC". Further inspection reveals that emails sent as one type or the other were almost always mistakenly recorded as both. In addition to all of these issues, no comprehensive or accurate list of job title labels is available.

2.3 Prior Analysis

The existing literature on analyzing social email behavior is mainly divided into two categories: traffic-based and social-based [14]. Traffic-based methods calculate statistics based only on email patterns. Social-based methods represent the email communications as a social graph and then extract information from this model using graph-based algorithms.

Using features extracted from email metadata, [15] was able to cluster levels of management at Enron. In addition to email traffic statistics, using features such as the presence of different email attachment types and the length of emails was shown to successfully categorize email behavior in [16].

Relational ties can be modeled as a graph network where nodes represent people and edges represent email interactions. This is a useful model because many statistics can be calculated from the layout of a social graph [17]. A common metric that has been shown to indicate importance in a social graph is betweenness centrality, which comes in several different flavors first developed by [18]. Betweenness centrality is a measure of how many shortest paths in a graph travel over each node. A node with high betweenness centrality in a social graph has been shown to represent a high degree of influence on other nodes. As shown in [19], a betweenness centrality algorithm can be used to determine community structures within an organization. Other successful metrics were used in [20], which detected the most important email users within a corporate network without using betweenness as a feature. The features used in that study were: degree, the number of edges connected to a node; density, the ratio of actual edges to the number of possible edges; and proximity prestige, the ratio of nodes that can reach a node i to the average distance from those nodes to i.

Instead of considering only traffic-based or social-based analytics, these can be be used jointly. An example of this approach [21], which combined features such as number of emails, response time, cliques, and degree centrality into a "Social Score", was used to rank Enron employees.

Chapter 3

Email Dataset and Feature Extraction

Over the past decade, the Enron dataset has been widely used to study email behaviors because it is one of the only datasets available comprised of real-world corporate emails. A list of ground truth job titles was compiled by [22]. However, there are known issues with these labels. Due to difficulties with the Enron dataset, as described in Section 2, this study uses a new dataset generated from volunteers in one of the university's centers.

3.1 Data Collection

Due to the inherent privacy concerns, researchers worked with the Internal Review Board (IRB) to approve a dataset which maintains participants' privacy. This dataset is meant to be representative of metadata which any company could use without violating the privacy of their employees. Special care was taken to protect the privacy of those involved in the study. During the collection

process, all subject and body text was hashed, and all email metadata was stored in a MySQL database using scripts without any researchers observing any email text. Furthermore, any identifying information has been omitted from this publication.

The center's email server offers the ability to both digitally sign and encrypt emails. Sending an email with either or both of these options appears as a special type of attachment in the raw email file. Therefore, during the data collection process, information about if an email was signed or encrypted was recorded. This process was part of the comprehensive parser script developed for extracting information from each email.

The following is the metadata parsed from each email:

- Destination and source email address
- Email time stamp
- Subject prefix (e.g., Re:, Fwd:)
- Hash of subject after removing prefix
- Hash of body text
- Length of subject in characters
- Length of body text in characters
- Number of attachments
  - Indicator if email was digitally signed
  - Indicator if email was encryped

Table 3.1: A comparison between the internal dataset and the Enron email corpus.

	Center	Enron
Time	11/2012-11/2015	1/2000-9/2002
Distinct Email Addresses	32,118	75,406
Participants	37	249
Distinct Emails	585,096	252,759

3.2 Dataset Description and Statistics

Table 3.1 compares statistics between this internal dataset and the Enron corpus. This internal database is more modern, contains more emails, and covers a longer time period. However, it is comprised of fewer people than the Enron dataset.

The center divides its employees into six main areas: directors, graduate students, operations, outreach, project management (PM), and research. A chart showing the distribution of classes is shown below in Figure 3.1

While the study includes only 37 volunteers, the email metadata from these volunteers identified 32 additional employees of the center. These additional employees were included in the study when ground truth for their job was available and when they were identified in at least 100 email metadata records. This provides 69 total employees in the study.

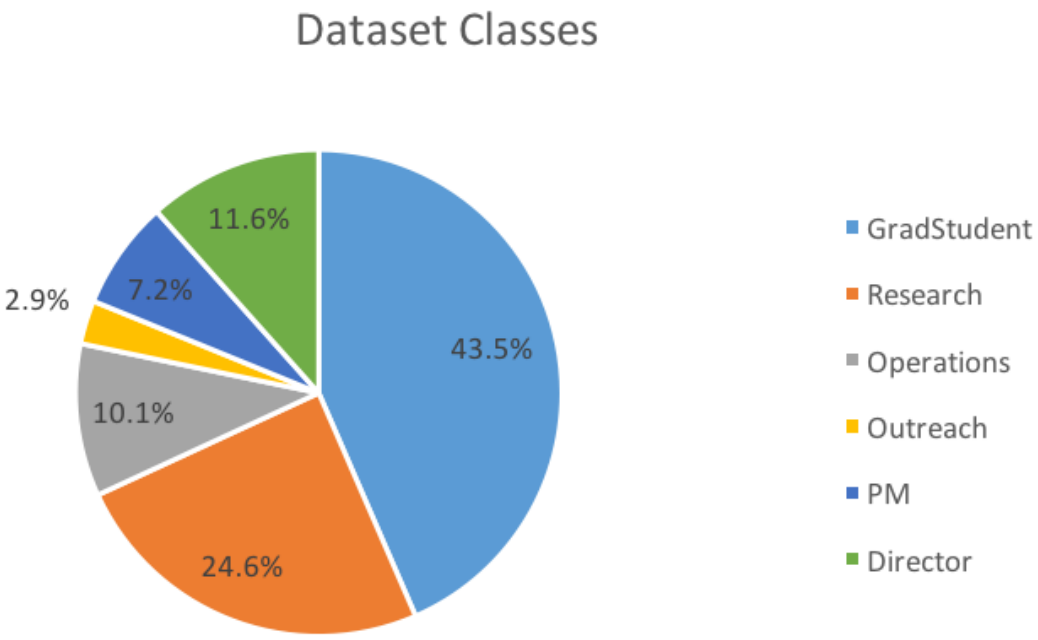

Figure 3.1: Pie chart showing the distribution of classes in the dataset. Note that the classes are far from being uniformly distributed.

3.3 Features

The study includes 111 features that were extracted from the email data: 80 traffic-based and 31 social-based. The traffic-based features are those calculated purely from the collected email metadata. The social features, on the other hand, first model the email patterns in a graphical network and then calculate statistics from this model. In the following sections, all features from each of the two categories are described.

3.3.1 Traffic-Based Features

Email Counts and Email Types

The simplest traffic-based features involve counting how many and what kinds of emails each participant sends and receives. First, the total number of emails, total sent, and total received give a measure of how active an email user is on average. These features can also indicate the direction tendencies of an employee's communication. Do they send more emails than they receive, or vice versa? All of the traffic-based features detailed below consider direction. Specifically, each metric is calculated three times: considering only sent emails, only received emails, and both sent and received emails.

Some traffic-based features focus on the different types of emails. Two examples of these types of features are the number of emails sent directly to each employee and the number of emails where they were carbon copied on the email. The opposite direction of this was inspected as well, that is, the number of emails the employees sent directly to others and the number of carbon copies sent out. The average number of recipients on emails sent and received for each participant were also calculated. Similarly, the number of emails sent and received as replies or forwards were used. These measures give a sense of how the employee communicates with others in the organization and their connectedness.

Several features were calculated from just the subset of emails that were digitally signed. These features were the total number of emails sent and received, number of unique email addresses, and the number of unique subjects. These same metrics were also calculated for encrypted emails.

Metadata Statistics

The metadata of the emails contains extremely useful information. This includes the email addresses involved, time stamp, the subject and body hashes and character counts, and the presence of any attachments. From the time stamp, the time of day for each email was available. The total number of emails with timestamps after hours were used as a metric. For this purpose, after hours was defined as between 6pm and 7am EST on weekdays or anytime on weekends. The timestamps were also used to calcualate the average number of emails per day for each employee. The mean and variance of the number of characters in the subject and body were calculated. The total number of attachments sent and received were computed as well as the average number of attachments per email.

Some of the most interesting information came from email addresses and subject hashes of the emails. The number of unique email address connections, both sent and received, was used as a feature. By counting the number of identical hashed subjects, it was determined how many unique subjects were both sent and received from each employee. The motivation behind these features is that employees with particular job titles may be more likely to be associated with long email chains, which would have the same subject. It was hypothesized that staff members had more external communications than graduate students. To test this, the number of emails sent and received from within the center and the university were calculated. Email addresses with a Virginia Tech domain were considered to be affiliated with the university. Email addresses with accounts on the internal mail server were labeled to be within the center. Note that all employee email

addresses of the center have a Virginia Tech domain, and are therefore also considered to be part of the university.

Most of the features described above involved raw email counts. However, this could skew data by giving more importance to employees who have been associated with the center longer. In order to normalize these values, corresponding percentage values were also fed into the learning algorithm. Examples include the percentage of sent emails that were sent after hours and the percentage of received emails with unique subjects out of all received emails.

Most Useful Traffic-Based Features

The best traffic-based feature for predicting employee status was the number of unique subjects received. The metric used to evaluate features is information gain. This was used because it is integral to the machine learning algorithm described in Section 4; details on the feature ranking analysis are provided in that section. This number of unique subjects received represents how many distinct conversations involve each individual. It is intuitive that people who are involved in more conversations hold a higher position in an organization because their input is requested more often. A histogram showing the different values for this metric over the different job classes is shown in Figure 3.2. It is clear from the figure that GradStudents participated in far fewer email conversations than any other group. Researchers generated more emails on average, but fewer than the Program Managers. The directors of the center participate in the most email conversations.

The second best traffic feature was the number of signed emails received. Signed emails usually

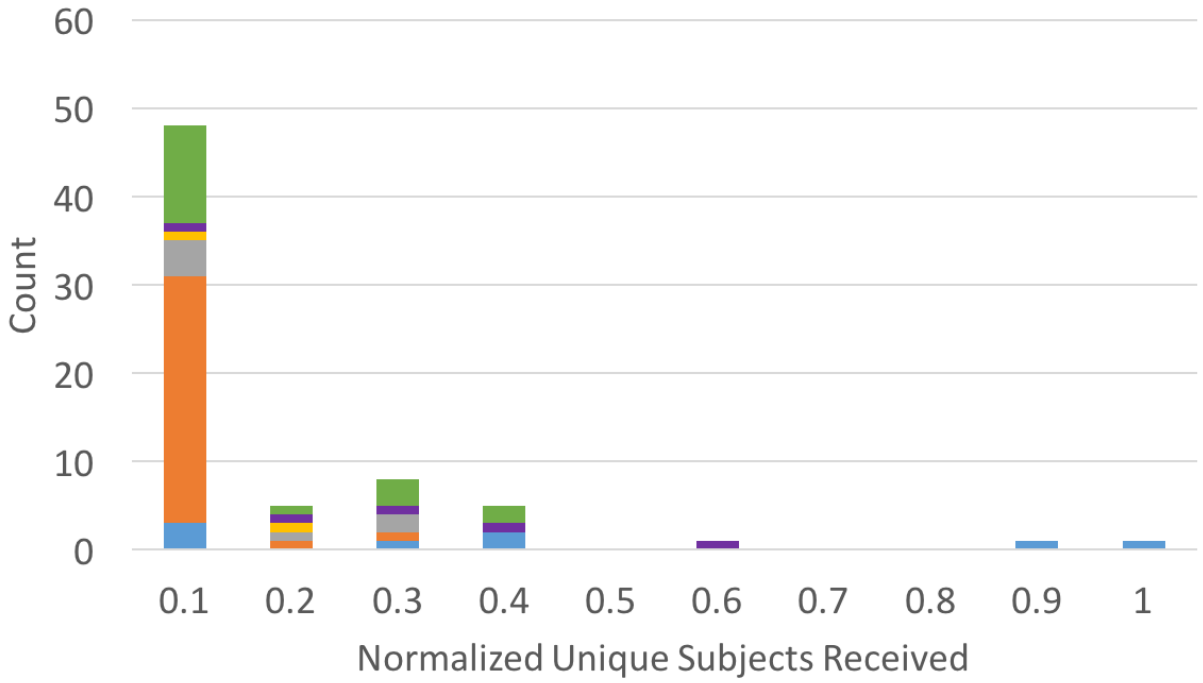

Figure 3.2: Histogram of unique subjects received by job title. The feature value for this plot has been normalized so that all values fall between 0 and 1. Note that by using different thresholds, meaningful splits in the data can be made. For example, all but two graduate students have a value less than 0.1.

signal sensitive information. Only certain groups within the center deal with this type of information, therefore it is understandable that this feature could help divide the subjects by title. Finally, the third best traffic feature was the number of emails received as forwards. Typically, those higher in the chain of command are forwarded emails where graduate students and lower-level employees are more likely to receive either replies or emails sent directly to them. Notice that there are intuitive explanations behind all of the features selected by the ranker.

3.3.2 Social Network Features

Social Network Representation

In addition to tracking metadata statistics, features are also derived from modeling the emails as a social network. A social network is composed of nodes, which represent people, and edges, which represent the emails between people. For this analysis, two different graphs were generated for analysis. In the full graph, an edge exists between any two individuals that exchanged at least one email. Each edge was given a weight equal to the total number of emails exchanged between the two employees. The edges are undirected for this analysis. A second graph only produces the same weighted edge between two nodes but only if at least 10 emails were exchanged. The purpose of this second graph is to filter out stray single-email relationships between coworkers that do not constitute meaningful communication. The full graph including only the study volunteers is plotted in Figure 3.3. Note that graduate students are found on the fringes of the network and are generally the least connected. However, there is a clear core group of employees that communicate very frequently.

Another representation of the full graph is shown as an adjacency matrix in Figure 3.4. Each of the two axes represent the employees of the center. The color at each coordinate indicates how much communication existed between the two employees. Some employees never exchanged any emails, while others exchanged many.

Using this model, another suite of statistics can be calculated about the people in the graph.

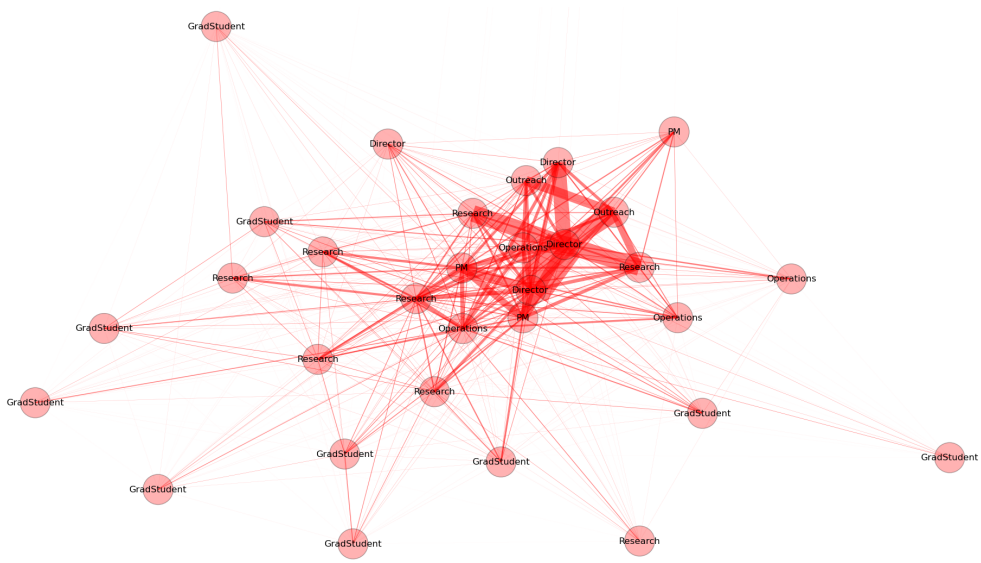

Figure 3.3: A social graph representation of the center. Nodes represent employees, and the thickness of the edges between nodes represent how many emails were exchanged.

Degree Measures

The degree of each node, from both the full and partial graph, was used as a feature. The degree of a node i is simply the number of other edges connected to node i. All graph features were calculated once for the full graph and again for the partial graph. The average neighbor degree was calculated used as another feature. The neighborhood of node i is comprised of all nodes that are connected to i via edges. Therefore for node i, this metric averages the degree of each node in the neighborhood of i. Mathematically, this is:

$$k_{\text{avg},i} = \frac{1}{|N(i)|} \sum_{j \in N(i)} k_j \tag{3.1}$$

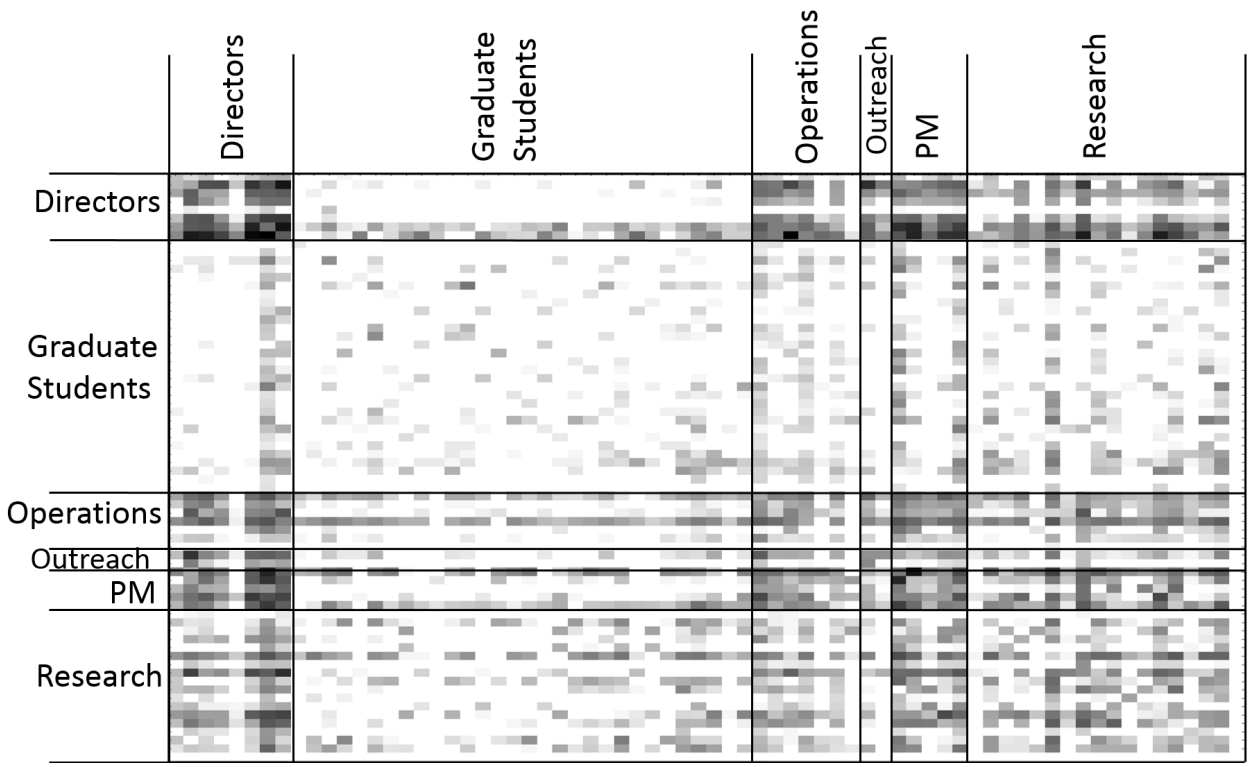

Figure 3.4: The adjacency matrix representing the social connections of the center. This graph is very well connected with just one component. Nonetheless, there are many pairs of individuals who never exchanged a single email.

where $N(i)$ are the neighbors of node i and k_j is the degree of node j. The distance between nodes were also used to generate some features. In graph theory, distance is measured by the length of the path between two nodes. Between node i and any other node j in graph $\mathcal{G}$, there exists a shortest path, $d(i, j)$. The average shortest path between node i and all other nodes in the graph, d_avg, i, was used as a metric for the learning algorithm. That is,

$$d_{avg,i} = \frac{1}{n-1} \sum_{j \in \mathcal{V}, j \neq i} d(i, j) \tag{3.2}$$

where n is the number of nodes in graph $\mathcal{G}$ and $\mathcal{V}$ is the set of nodes in $\mathcal{G}$. Similarly, the maximum shortest path length, or eccentricity, was used as a feature in the learning algorithm. All of these measures can be interpreted to represent the centrality of a node. If a node has many neighbors with large degrees or if it has very short maximum shortest paths it is probably representative of a person well-connected within the center.

Cliques

Some of the social features were based on existing graph theory concepts and algorithms. For example, cliques. If a subgraph of a graph $\mathcal{G}$ is maximally connected, that is all nodes are connected directly to each other, then this is called a maximal clique. The number of cliques to which a node belongs was used as a feature. The motivation behind using this metric is that it should mirror working groups within the center. Therefore, the more groups an employee belongs to or communicates with, the more important they are assumed to be.

Adapting Search Engine Algorithms

The hubs and authorities of each node in both graphs were calculated. The terms hubs and authorities come from the Hyperlink-Induced Topic Search (HITS algorithm) developed by [23]. This algorithm was originally designed to rate web pages, but has since been applied to social networks. A node's authority is just that—a measure of its importance over other nodes. A node's hub score is a measure of how well-connected it is to other nodes.

Another algorithm used to generate features was the pagerank algorithm, developed by Google [24] also to rank webpages for search results. The assumption is that the most important webpages will be linked to frequently by other pages. Therefore, the ranking is determined by estimating the quality and quantity of links to a node. Both of these algorithms were shown to predict expertise within an online social network in [25].

Clustering Metrics

The triangle clustering coefficient, C_3 was also used as a metric. Say there exists a node i with neighbors m and n. This value, developed by [26], measures the probability that m and n are also connected. This is calculated by comparing the number of triangles within the graph to the maximum number of possible triangles in the graph. If node i has degree k_i, there can be at most $\frac{k_i(k_i-1)}{2}$ triangles formed in this subgraph. Recall that the social networks are weighted graphs. The weights of the graph are incorporated to this metric by finding the geometric mean [25]. Therefore, triangle clustering coefficient for node i, $C_{3,i}$, is:

$$C_{3,i} = \frac{2}{k_i(k_i-1)} \sum_{m,n} (\tilde{w}_{i,m}\tilde{w}_{m,n}\tilde{w}_{n,i})^{\frac{1}{3}} \tag{3.3}$$

The edge weights in this calculation must be normalized compared to the maximum weight in the subgraph, i.e. $\tilde{w}_{i,m} = \frac{w_{i,m}}{max(w_{i,m})}$.

The square clustering coefficient is very similar and was also used in the algorithm. This metric, developed by [27], measures the probability that m and n are also neighbors to a fourth node, p. This configuration would form a square. To simplify the calculations, the graph for this metric is

viewed without edge weights. Therefore, the square clustering coefficient for node i, $C_{4,i}$, is the proportion of actual squares within a subgraph centered around node i to the maximum number of possible squares in the same subgraph. This is calculated as:

$$C_{4,i} = \frac{\sum_{m=1}^{k_i} \sum_{n=m+1}^{k_i} q_i(m,n)}{\sum_{m=1}^{k_i} \sum_{n=m+1}^{k_i} [a_i(m,n) + q_i(m,n)]} \tag{3.4}$$

where $q_i(m,n)$ is the number of neighbors shared by m and n, excluding i and

$$a_i(m,n) = (k_m - \eta_i(m,n))(k_n - \eta_i(m_n)) \tag{3.5}$$

where $\eta_i(m,n) = 1 + q_i(m,n) + \theta_{mn}$. It is defined that θ_{mn} is an indicator function that takes on a value of 1 if m and n are neighbors and 0 otherwise. In theory, the higher the clustering coefficient, the more connected the node is within its neighborhood.

Centrality Measures

The majority of the social-based features were variations on centrality measures. First is closeness centrality. Closeness centrality, $C(i)$ is the normalized inverse of the sum of shortest path distances from node i to all other nodes in the graph [28]. It is calculated as follows:

$$C(i) = \frac{n-1}{\sum_{j=1}^{n-1} d(i,j)} \tag{3.6}$$

where n is the number of nodes in graph $\mathcal{G}$ and $d(i,j)$ is the minimum shortest path distance between node i and node j.

Next is betweenness centrality. In a graph, there exists a shortest path between any node s and any other node t. Betweenness centrality of a node i, $C_B(i)$, is the percentage of all shortest paths in

graph $\mathcal{G}$ that traverse node i [18]. It is calculated:

$$C_B(i) = \sum_{s,t \in \mathcal{V}} \frac{\sigma(s,t|i}{\sigma(s,t)} \tag{3.7}$$

where $\mathcal{V}$ is the set of all nodes in $\mathcal{G}$, $\sigma(s,t)$ is the number of shortest paths between s and t, and $\sigma(s,t|i)$ is the number of those paths that pass through i. It is further defined that if $s = t$, then $\sigma(s,t) = 1$ and if $i \in s,t$, then $\sigma(s,t|i) = 0$.

Degree centrality of a node i, $C_{d,i}$ is simply the percentage of nodes within the graph that are connected to node i [29]:

$$C_{d,i} = \frac{k_i}{(n-1)} \tag{3.8}$$

where k_i is the degree of node i and n is the number of nodes in graph $\mathcal{G}$.

Current flow betweenness centrality, also known as information centrality, is measured for each node. In general, metrics related to betweenness centrality differs from the previous centrality measures in that they consider all paths between nodes instead of exclusively shortest paths [**?**]. Current flow betweenness centrality in particular was modeled after how current flows in electrical networks. In circuits, current is distributed over the possible paths; in this metric a similar proach is taken to determine the information content of each path between two nodes. For a graph $\mathcal{G}$ where all nodes are reachable, it is possible to construct a matrix B such that:

$$b_{ii} = 1 + \text{sum of weights of all edges connected to node } i \tag{3.9}$$

$$b_{ij} = 1 - w_{ij} \tag{3.10}$$

where w_{ij} is the weight of the edge connecting nodes i and j. If i and j are not neighbors, w_{ij} is

defined to be 0. Denote $B^{-1} = C$. From this matrix, the current flow betweenness centrality of node i, I_i, is calculated as:

$$I_i = \frac{n}{nc_{ii} + T - 2R} \tag{3.11}$$

where n is the number of nodes in $\mathcal{G}$, T is the sum of the diagonal elements, $T = \sum_{j=1}^{n} c_{jj}$, and R is the sum of any row in C, $R = \sum_{j=1}^{n} c_{ij}$. Note that because the matrix is symmetric, all rows of C will have the same sum, therefore R is not dependent on i.

This includes current flow closeness centrality, current flow betweenness centrality [30], communicability centrality, communicability betweenness centrality [31], and load centrality [32].

All of these different graph statistics were used as inputs into the random forest algorithm to characterize each node's importance in the social graph.

Most Useful Social Network Features

The histogram of hub values in the full graph broken down by class is shown in Figure 3.5.

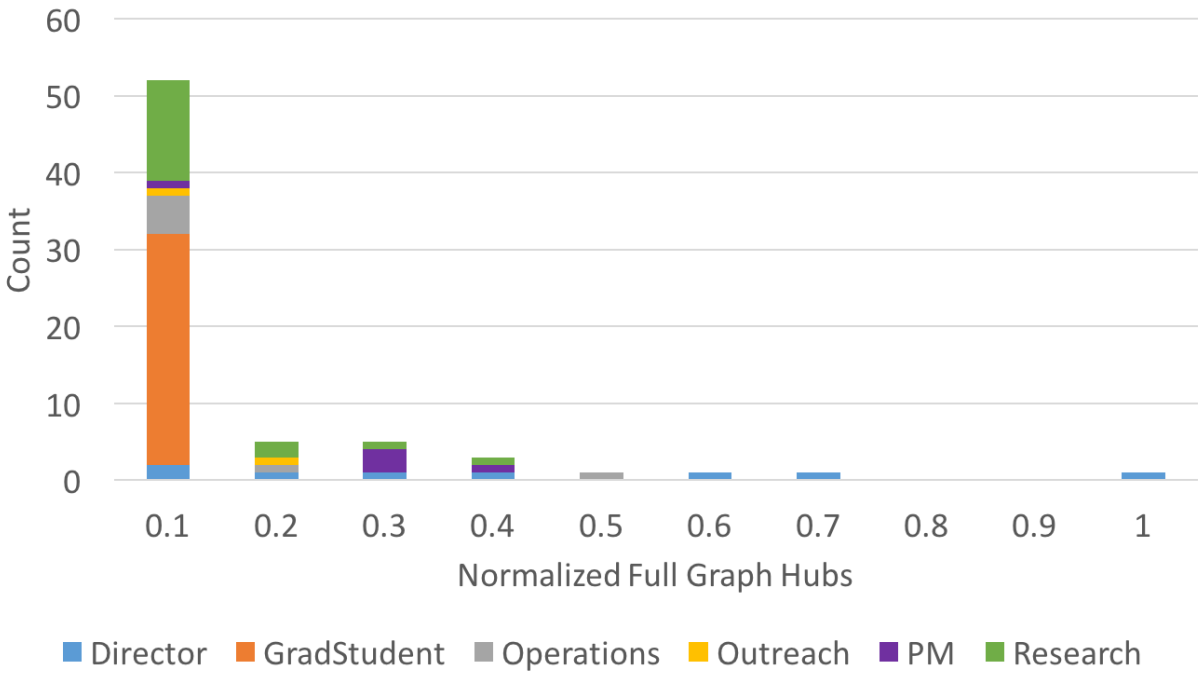

Figure 3.5: Histogram of hubs from social graph by job title. Note that directors on average have the highest hub score and graduate students have the lowest. In fact, three out of eight directors can be identified by filtering samples on full graph hubs values > 0.6.

Chapter 4

Algorithm Design

4.1 Tools

4.2 Model Selection

Due to the large number of features and relatively low number of participants, a classification method was carefully chosen to avoid overfitting the data. While tree based classifiers can be susceptible to overfitting, the random forest is robust to overfitting issues and was therefore chosen for this study. The java-based software package Weka was used to generate the random forest based on the algorithm described in [33].

Random forests are an ensemble method of machine learning, comprised of many random trees. A random tree is a machine learning algorithm that uses training data to learn a series of rules

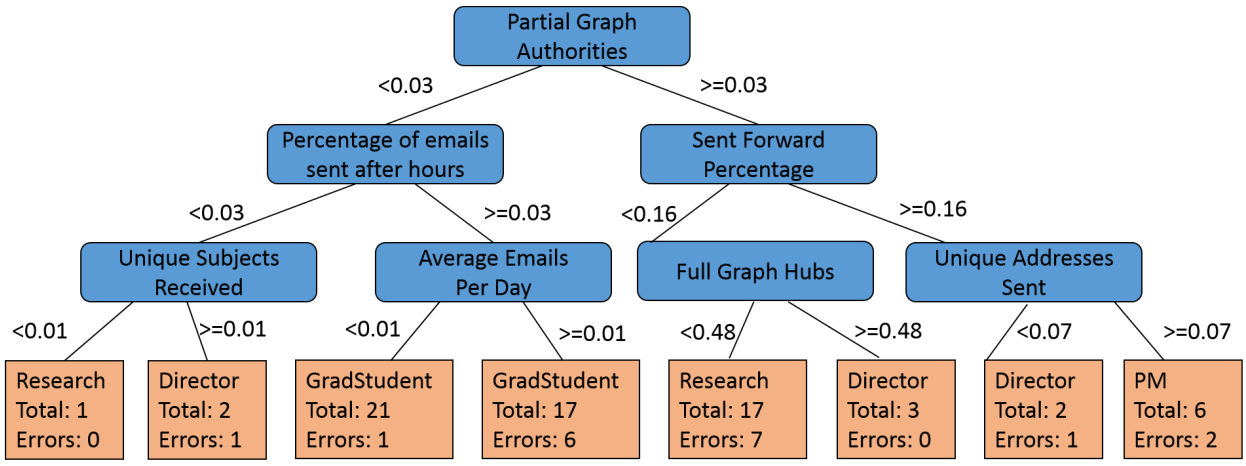

Figure 4.1: Example random tree of depth 3 to demonstrate how a few rules can be used to find significant class divisions.

for classification. These rules are constructed in a hierarchy that visually resembles a tree. Each decision is based on what rule will maximize the information gained. An example random tree with depth three is shown in Figure 4.1.

Random forests build deep random trees with slight random variations. Individually, these random trees overfit the data. However, these random trees are combined through a process of bootstrap aggregating, or bagging. The number of trees was chosen to be 750 based empirical analysis. The bagging process involves each random tree generating a new training data set by sampling observations from the input training set with replacement. These subsamples are used to build the random trees. For this analysis, each tree selects $\frac{2N}{3}$ samples to train the trees where N is the number of data points in the overall training set. Just as the samples were subsampled, so were the features. Only this subset of features can be used as rules for that tree. In this implementation, each tree used a subsample of 15 random features.

After all the trees are built, the test data is run through all the random trees in the forest. Each tree outputs a prediction label for each data point, and the majority vote on each sample is the final predicted label. This random forest model reduces the variance and increases the accuracy of the model compared to a single random tree.

4.3 Feature Selection

Random forests can be difficult to interpret because the ensemble method obscures which features are most meaningful. An attribute analysis helps to better understand which features are better label predictors. Since random trees use information gain to dictate splits, information gain was used as the evaluation criteria for the features. Specifically, each attribute was evaluated by measuring the information gain with respect to the class. Information gain is calculated as follows:

$$I(\text{Class}; \text{Attribute}) = H(\text{Class}) - H(\text{Class}|\text{Attribute}) \tag{4.1}$$

where $I(\text{Class}; \text{Attribute})$ represents the mutual information between the class and the attribute, $H(Class)$ is the entropy of the class variable, and $H(\text{Class}|\text{Attribute})$ represents the conditional entropy of the Class given the Attribute value.

Mutual information represents how well knowledge of the attribute informs the prediction of the class. In this model, both the attribute and the class are treated as random variables. The entropy of a random variable is a measure of the uncertainty associated with it. After this information gained value was calculated for each feature, they were ranked in order of most important to least. Table 4.1 shows the top twenty features from this analysis and the features' corresponding

information gain.

Table 4.1: Top 20 features ranked by the information gain.

Feature	Type	Ranker
Unique subjects received	Traffic	0.728
Total signed emails received	Traffic	0.728
Number of emails received as forwards	Traffic	0.719
Full graph hubs	Graph	0.589
Partial graph communicability centrality	Graph	0.554
PG communicability betweenness centrality	Graph	0.554
Number of emails received as CC	Traffic	0.507
Percentage of emails received as forwards	Traffic	0.503
Partial graph degree centrality	Graph	0.492
Partial graph pagerank	Graph	0.492
PG current flow closeness centrality	Graph	0.492
Average number of emails received per day	Traffic	0.489
Average total emails per day	Traffic	0.479
Partial graph average shortest paths	Graph	0.476
Partial graph closeness centrality	Graph	0.476
Unique email addresses from signed emails	Traffic	0.457
Number of emails sent as CC	Traffic	0.43
Number of emails received as replies	Traffic	0.43
Average emails sent per day	Traffic	0.404

Chapter 5

Performance Analysis

The results section first shows the algorithm's ability to correctly classify both the study's volunteers and the additional employees identified from the volunteers' emails. The second part of the results section assumes perfect labeling of the employees and analyzes interactions between employees of different job titles. The ultimate goal of this research was to determine what additional information can be gained by analyzing the organic organizational chart when compared with the official organizational chart.

5.1 Classification Results

Data was split by randomly assigning each email to either the training or testing set with equal probability. Then, all of the metrics described in Section 3 were calculated for both groups separately. The training data was used as input to the random forest algorithm as described in Section 4,

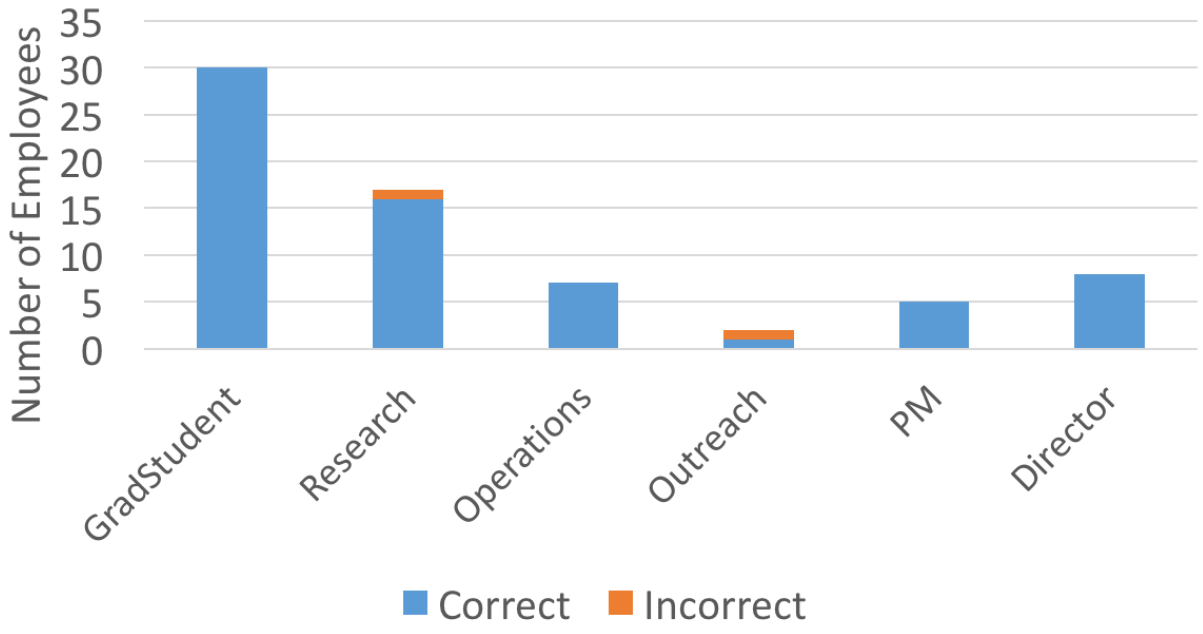

Figure 5.1: The Random Forest algorithm was extremely accurate even for very uneven class sizes. Note that all members of 4 classes were labelled perfectly. There were only 2 errors out of 69 employees, both of which for employees who did not provide emails for the study.

and predictions were generated for the test data. The number of correct and incorrect classifications for each class are shown below in Figure 5.1. Note that only two predictions were wrong: one person each in research and outreach were misclassified as graduate students. It is important to note that both misclassifications are for employees who did not provide their emails for the study. Therefore, the classification accuracy for the study participants is 100%, correctly classified inferred employees is 93.75%, and the overall accuracy of this method using all features is 97.1%.

Note that this method relies on some assumptions. One is that employees with the same title exhibit similar email behavior. Overall, based on the success of the algorithm and the distributions of the histograms, this seems to prove true. Another premise underlying this analysis is that peoples'

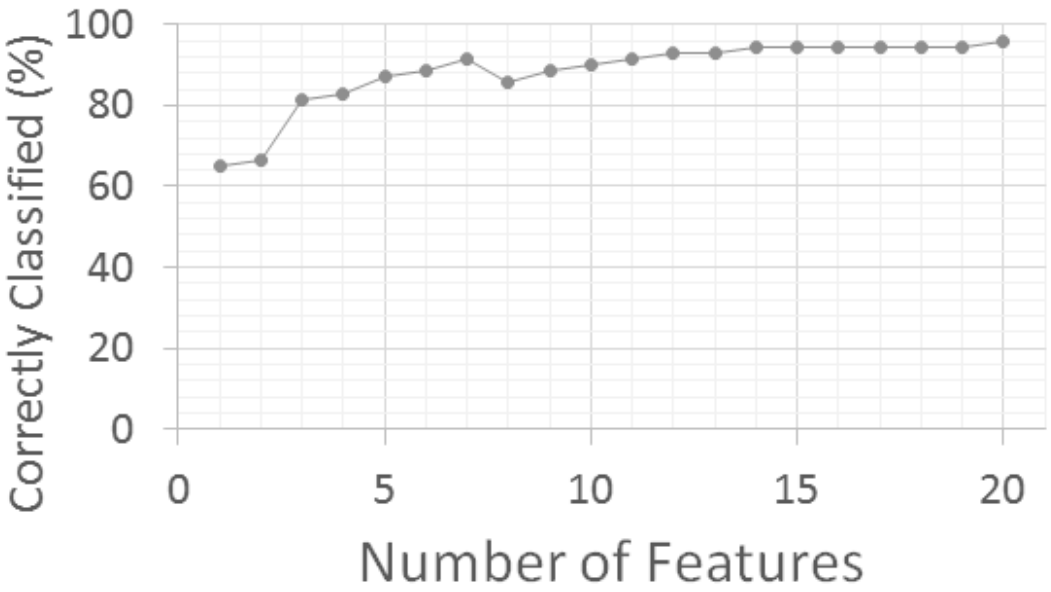

Figure 5.2: Prediction accuracy compared to number of features used for analysis. Note that the accuracy is still very high, 95.6%, when only twenty features are used. The outcome of using only the top twenty features produces three classification errors, only one more than using the full set of 98 features.

email behaviors are consistent over time. This seems to be true for the time range in this study.

To determine which features were necessary to the analysis, the algorithm was run several times with a subset of the features. The first subset used only the top 20 features from Table 4.1. Using only these features resulted in 3 classification errors, or 95.6% accuracy. This is only one more error than was found using all 98 features. For each subsequent run, the least useful feature according to the feature analysis was removed from the input to the system until only one feature remained. A plot of this analysis is shown below in Figure 5.2. Even using just the top thee features resulted in classification accuracy over 80%. Therefore, a very good classifier can be built using much fewer features if the features are selected properly.

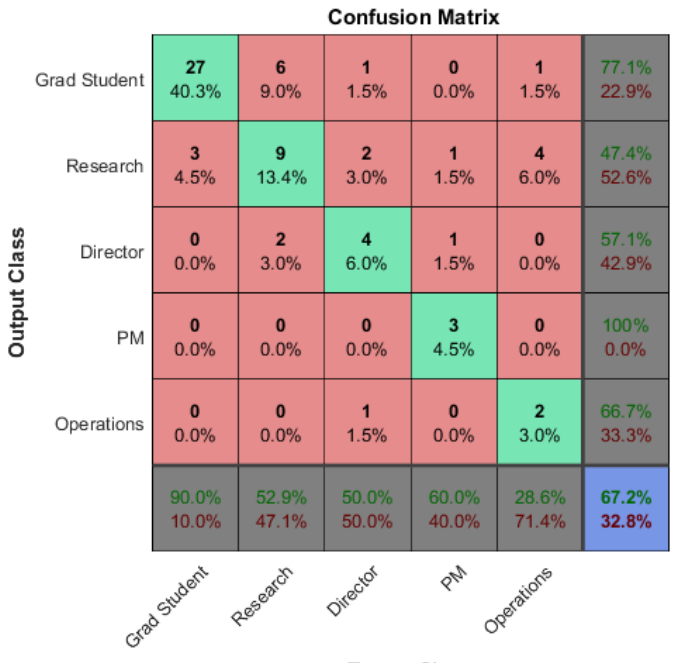

Figure 5.3: Prediction accuracy was higher for graduate students, who have more uniform behaviors and more training data. However, for each category, the majority of the people classified were true members of that group, save for research. Note that even for those classified as research, there were more true researchers than any other class.

5.2 Leave-One Out Cross Validation

So I tried LOOCV which gave, 5.3.

5.3 Hierarchy Analysis

Most of the employees at the center are organized under a director and work with a program manager (unless for example they are a director or program manager). To generate a metric of

how well emails can be used to predict the center's organizational chart, the director and project manager for each applicable employee is predicted from the email metadata.

The director of each employee is predicted by the algorithm to be the director that the employee communicated with most by email. Only 57.58% of the center's employees communicate most frequently with their official director. This result points to a possible disconnect between the official organization chart and the organic relationships within the center.

To identify each employee's project manager ground truth is selected to be the project that primarily funds the employee. This time, 72.73% of graduate students and researchers communicate most frequently with their primary program manager. The relation between employees to project managers appears to be stronger than that with directors. Many of the errors in this classification are due to employees who work with multiple project managers.

Chapter 6

Future Work

6.1 Generalization

- As discussed above, the accuracy of the leave-one-out cross-validation test was low because of the limited number of data points. The confusion matrix in Figure 5.3 shows that about one third of the predictions were incorrect.
- These results can typically be improved by increasing the amount of training data.
- Performed experiment to gauge how the amount of training data affects the accuracy of the LOOCV test. This is shown below in Figure 6.1.

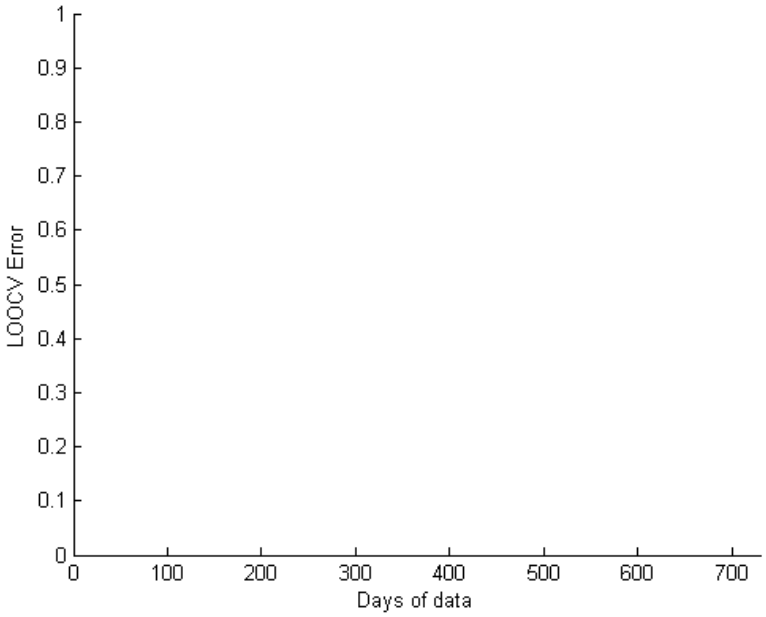

Figure 6.1: Effects of using more data for leave-one-out cross-validation. [Still working on generating this, but I expect it to exponentially decrease. From this data, I can fit a curve to the data that I do have to estimate how well the algorithm will perform with more data.]

6.2 Deep Learning

- Techniques such as random forests work well for small, low-dimensional data. However, imagine if the size of the dataset were greatly increased to billions of people without labels. In this case, a more sophisticated model would be necessary for successful classification. This is required to accurately learn the complicated structure inherent to such large real-world datasets. Suppose a representative sample of volunteers from the dataset self-identifies themselves. This would produce a set of labeled training data, transforming the problem space to semi-supervised learning.

- Describe deep belief networks: A deep belief network is a generative graph with nodes

that represent stochastic variables. There are several layers of these nodes, some of which are visible and some are hidden. Typically, the top two layers of nodes have undirected connections. Therefore, these layers are actually Restricted Boltzman machines (RBMs). The rest of the layers use directed edges, forming a Bayesian network. For each node, there exists a probability of activation, $p(s_i = 1)$, which is represented by the nonlinear sigmoid function applied to a weighted input from the layer above. Specifically,

$$p(s_i = 1) = \frac{1}{1 + \exp\left(-b_i - \sum_j s_j w_{ij}\right)} \tag{6.1}$$

where j represents the ancestors of node i, w_{ij} are the weights on the connections between i and j, and b_i is the bias associated with node i...

- The challenge with very deep networks is that they are very difficult to train [34]. Unsupervised pre-training can improve the efficency of training deep belief networks, as first described in [35]. In that work, Hinton describes a fast, greedy method to train these networks one layer at a time.

- Describe the process of unsupervised pre-training.

- Testing on the MNIST dataset of handwritten digits, unsupervised pre-training is shown to improve classification results over only supervised training [36]. Pre-training identifies a set of initial weights for the network that can ultimately lead to better classification results. An advantage of unsupervised pre-training is that, with small enough layers, it acts as a regularizer by decreasing variance and increasing the bias [35].

- This labeled data can be used for backpropagation to fine-tune the network. Describe the process backpropagation with the small set of labeled data.

- In conjunction with greedy pre-training, backpropagation has been shown to improve both the optimization and generalization of a DBN. Hinton et al. showed that unsupervised pre-training followed by backpropagation of a DBN outperformed traditional feed forward neural networks on the MNIST dataset [37].

- Analogous paper for comparison:

  Much of the research in deep learning today is focused on image classification and analysis, such as the MNIST dataset. However, DBNs have been applied to text analysis as well. In [38], a sophisticated version of a DBN was used to construct a sentiment classifier. The purpose of this research is to label online reviews as positive, negative, or neutral. The difference between this method and other DBN approaches is that this work replaced some of the hidden layers of the network with a feature selection step. Overall, this improves the training learning efficiency of the algorithm. This method can classify sentiments more accurately and train more quickly than previous semi-supervised algorithms.

Chapter 7

Conclusions

This work presents a new dataset, approximately the size of Enron, that was collected from volunteers' emails with particular attention to protect participant privacy. The new dataset includes accurate labels executed by researchers with knowledge of the center and its employees. Statistics were calculated from this dataset, and were used with a random forest algorithm to automatically classify the center's employees. Random Forests are shown to be powerful classifiers by predicting employee job titles with very high accuracy, even for employees for whom only secondhand data is available in the dataset. Using only 3 features, employees are successfully classified higher than 80% of the time and are classified over 95% of the time when 20 features are used. The email data was also used to show that emails could be used to predict an employee's primary program manager, but had a worse chance of being able to identify the director associated with the employee on the official organizational chart. This work has shown that it is possible to generate important organizational information from using carefully processed email metadata without compromising the

privacy of employees. Future work will attempt to gain more insight into an organization's organic hierarchy and to apply these algorithms to other datasets to determine the general applicability of the results.

Bibliography

[1] S. Radicati and Levenstein, *Email statistics report, 2015-2019*. Technical report, 2015.

[2] D. Fisher, A. J. Brush, E. Gleave, and M. A. Smith, "Revisiting Whittaker & Sidner's email overload ten years later," in *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pp. 309–312, ACM, 2006.

[3] B. Klimt and Y. Yang, "Introducing the Enron Corpus.," in *CEAS*, 2004.

[4] E. M. Bahgat, S. Rady, and W. Gad, "An E-mail Filtering Approach Using Classification Techniques," in *AISI2015, November 28-30, 2015, Beni Suef, Egypt*, vol. 407, pp. 321–331, Cham: Springer International Publishing, 2016.

[5] R. Shams and R. Mercer, "Classifying Spam Emails Using Text and Readability Features," in *ICDM 2013*, pp. 657–666, Dec. 2013.

[6] B. He, Z. Li, and N. Yang, "A Novel Approach for Email Clustering Based on Semantics," in *WISA, 2014*, pp. 269–272, Sept. 2014.

[7] P. S. Keila and D. B. Skillicorn, "Structure in the Enron email dataset," *Computational & Mathematical Organization Theory*, vol. 11, no. 3, pp. 183–199, 2005.

[8] Z. Sofershtein and S. Cohen, "Predicting Email Recipients," pp. 761–764, ACM Press, 2015.

[9] Q. Hu, S. Bao, J. Xu, W. Zhou, M. Li, and H. Huang, "Towards building effective email recipient recommendation service," in *SOLI 2012*, pp. 398–403, July 2012.

[10] A. Nordb, "Data Visualization for Discovery of Digital Evidence in Email," 2014.

[11] K. K. Waterman and P. J. Bruening, "Big Data analytics: risks and responsibilities," *International Data Privacy Law*, vol. 4, pp. 89–95, May 2014.

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

[13] J. Shetty and J. Adibi, "The Enron email dataset database schema and brief statistical report," *Information sciences institute technical report, University of Southern California*, vol. 4, 2004.

[14] G. Tang, J. Pei, and W.-S. Luk, "Email mining: tasks, common techniques, and tools," *Knowledge and Information Systems*, vol. 41, pp. 1–31, June 2013.

[15] K. Yelupula and S. Ramaswamy, "Social network analysis for email classification," in *Proceedings of the 46th Annual Southeast Regional Conference on XX*, pp. 469–474, ACM, 2008.

[16] S. Martin, B. Nelson, A. Sewani, K. Chen, and A. D. Joseph, "Analyzing Behavioral Features for Email Classification.," in *CEAS*, 2005.

[17] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge University Press, Nov. 1994.

[18] L. C. Freeman, "A Set of Measures of Centrality Based on Betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[19] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, "Email as Spectroscopy: Automated Discovery of Community Structure within Organizations," *arXiv:cond-mat/0303264*, Mar. 2003. arXiv: cond-mat/0303264.

[20] G. Wilson and W. Banzhaf, "Discovery of email communication networks from the Enron corpus with a genetic algorithm using social network analysis," in *CEC'09*, pp. 3256–3263, IEEE, 2009.

[21] R. Rowe, G. Creamer, S. Hershkop, and S. J. Stolfo, "Automated social hierarchy detection through email network analysis," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 109–117, ACM, 2007.

[22] J. Shetty and J. Adibi, "Ex employee status report.," 2004. [http://www.isi.edu/~adibi/Enron/Enron_Employee_Status.xls]. _Internet Archive_.[https://web.archive.org/web/20131126121206/http://www.isi.edu/~adibi/Enron/Enron_Employee_Status.xls], Accessed 1/30/2016.

[23] J. M. Kleinberg, "Hubs, authorities, and communities," *ACM Computing Surveys (CSUR)*, vol. 31, no. 4es, p. 5, 1999.

[24] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: bringing order to the Web.," 1999.

[25] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: structure and algorithms," in *Proceedings of the 16th international conference on World Wide Web*, pp. 221–230, ACM, 2007.

[26] J. Saramki, M. Kivel, J.-P. Onnela, K. Kaski, and J. Kertsz, "Generalizations of the clustering coefficient to weighted complex networks," *Physical Review E*, vol. 75, Feb. 2007.

[27] P. G. Lind, M. C. Gonzlez, and H. J. Herrmann, "Cycles and clustering in bipartite networks," *Physical Review E*, vol. 72, Nov. 2005.

[28] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1978.

[29] S. P. Borgatti and D. S. Halgin, "Analyzing affiliation networks," *The Sage handbook of social network analysis*, pp. 417–433, 2011.

[30] U. Brandes and D. Fleischer, *Centrality measures based on current flow*. Springer, 2005.

[31] E. Estrada and N. Hatano, "Communicability in complex networks," *Physical Review E*, vol. 77, no. 3, p. 036111, 2008.

[32] M. E. Newman, "Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality," *Physical review E*, vol. 64, no. 1, p. 016132, 2001.

[33] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[34] Y. Bengio, "Learning Deep Architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[35] G. E. Hinton, "To recognize shapes, first learn to generate images," *Progress in brain research*, vol. 165, pp. 535–547, 2007.

[36] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?," *The Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.

[37] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[38] P. Ruangkanokmas, T. Achalakul, and K. Akkarajitsakul, "Deep Belief Networks with Feature Selection for Sentiment Classification,"