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

A. A. (Louis) Beex

R. Michael Buehrer

Bert Huang

March 25, 2016

Blacksburg, Virginia

Keywords: Data analytics, machine learning, social computing

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

Kayla M. Straub

(ABSTRACT)

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. 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 thesis demonstrates a suite of methods to an appropriately anonymized academic email dataset created from volunteers' email metadata. This new dataset, from an internal email server, is first used to validate feature extraction and machine learning algorithms in order to generate insight into the interactions within the center. Based solely on email metadata, a random forest approach models behavior patterns and predicts employee job titles with 96% accuracy. This result represents classifier performance 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 center's formal operating structure. The culmination of this work is an organic organizational chart, which contains a fuller understanding of the center's internal structure than can be found in the official organizational chart.

Acknowledgments

I want to acknowledge the many people who helped me with my graduate work at Virginia Tech and made this work possible. Thank you to the members of my committee: Dr. Robert W. McGwier, Dr. Louis Beex, Dr. Michael Buehrer, and Dr. Bert Huang for your guidance. In particular, I would like to sincerely thank Dr. McGwier for his vision and support throughout this research.

I want to express my gratitude to the faculty, staff, and students of the Hume Center for all of your help and for providing me with the opportunities to work on such interesting research. I am especially thankful for Dr. Joseph Ernst and Dr. William C. Headley for their invaluable insight and assistance.

I would also like to thank my family and friends for their support throughout my college career. I want to specifically thank my husband Ben for his unwavering support and for always pushing me to be my best.

Finally, I thank God for all the blessings in my life and for getting me to where I am today.

Contents

List of Figures **vi**

List of Tables **viii**

1 Introduction **1**

1.1 Motivation 1

1.2 Problem Overview 3

1.3 Approach 4

1.4 Contributions 4

1.5 Outline 5

2 Prior Work **6**

2.1 Background 7

2.2 Commonly Used Datasets 7

2.3 Prior Email Analysis 9

2.3.1 Machine Learning Overview 11

3 Email Dataset and Feature Extraction **12**

3.1 Data Collection 12

3.2 Dataset Description and Statistics 15

3.3 Features 19
3.3.1 Traffic-Based Features 19
3.3.2 Social Network Features 25

4 Algorithm Design **38**

4.1 Algorithm Selection 38
4.2 Algorithm Description 40
4.3 Feature Selection 45

5 Performance Analysis **47**

5.1 Classification Results 48
5.2 Leave-One-Out Cross-Validation 52
5.3 Hierarchy Analysis 55

6 Future Work **59**

6.1 Improving LOOCV Accuracy 60
6.2 Algorithms for Processing Larger Data 62
6.2.1 Deep Belief Networks 63
6.2.2 Collective Classification 66

7 Conclusions **67**

8 Bibliography **69**

List of Figures

3.1 Example subject hash 13

3.2 Center emails over time 17

3.3 Dataset class distribution 18

3.4 Emails per person per job title 19

3.5 Forwarded emails received histogram 24

3.6 The social network of the center 26

3.7 The dataset represented as an adjacency matrix 27

3.8 Types of emails employees sent to themselves 28

3.9 Histogram of the hubs feature for the partial graph 36

4.1 Algorithm selection analysis 39

4.2 Example random tree 40

4.3 Random forest prediction process 43

5.1 Job title classification results 49

5.2 Job title prediction confusion matrix 50

5.3 Accuracy of classification over time 51

5.4 Prediction accuracy compared to number of features 52

5.5 Leave-one-out cross-validation results 53

5.6 Generated organic organization chart of the center 56

5.7 Project manager and director prediction results 58

6.1 Effects of more data on prediction accuracy 61

6.2 Effects of more people on prediction accuracy 62

6.3 The Components of a Deep Belief Network 63

List of Tables

3.1 A comparison between the internal dataset and the Enron email corpus. 16

4.1 Top 20 features ranked by the mutual information. 46

Chapter 1

Introduction

1.1 Motivation

A reorganization of a business can be very costly and has far-reaching effects once implemented, either for better or for worse. Some of the world's largest corporations have recently executed some major restructuring. For example, in the last six months, Northrop Grumman consolidated two of its four business sectors: electronic systems and information systems. These changes also included creating a new executive position filled by the former head of electronic systems [1]. Another major merger is the deal struck by two major chemical corporations: DuPont and Dow. They agreed to combine their resources and then split off into three new entities. This is a supremely complex task involving the reassignment of over $100,000$ people [2]. All of these companies use

official organizational charts to manage the corporate hierarchy. Reorganizing the structure of the corporation involves a reorganization of this official chart.

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 could be invaluable during such a business reorganization, 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.

Extracting hierarchical relationships from email metadata can be used for several potential applications beyond gaining a better understanding of corporate operations. These methods could be applied to any communications system such as cellular communications, website links, social media, and network traffic. Within these sytems it would be possible to perform organizational analysis, anomalous behavior detection, or leadership identification. For example, it could be possible to identify people of influence or potential leaders based on their volume and patterns of communication. Alternatively, by recognizing behavior patterns inconsistent with a person's surrounding peers, potential insider threats could be identified. In a military context, this work shows the importance of protecting this metadata from adversaries who could use it to uncover information about an organization that they could use to find vulnerabilities. This type of work is also very important to government agencies in order to analyze large-scale data, as was made clear by the reports released by Edward Snowden in 2013 [3]. The type of analysis described in this

work applied to the data Snowden reports about could produce valuable understanding of people's behavior and status.

1.2 Problem Overview

This thesis addresses the problem of automatic organizational determination. Specifically, it considers using email analysis to perform this operation. However, there is a lack of modern, publicly available email datasets to use for this research. Out of the email datasets that do exist, namely the Enron email corpus, there is a lack of accurate job title labels for the employees. Labels are necessary for evaluating the results of the organic hierarchy analysis with the official organization chart.

The research presented in this thesis documents how a new email metadata dataset was collected and appropriately anonymized from the email records of 37 voluntary participants from an academic research organization. This metadata is used not only to analyze the participants, but also to what extent the non-participant members of the organization can be characterized.

The problem of analysis is broken into two parts. The first half addresses how to accurately classify employees by their job titles. This will give insight into how well email behaviors indicate the official hierarchy as well as how similar the job titles are to one another. The second part of the problem investigates how well this organic organizational chart matches the official organizational

chart. Altogether, the result reveals how much information about an organization can be extracted from the internal emails.

1.3 Approach

These emails are used to calculate a high-dimensional feature set to be used with machine learning techniques. The 114 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 accurately predict each employee's job title. Finally, a comparison is presented between relationships displayed in the data with the formal organizational chart. Combining predicted job titles with hierarchical relationships discovered from the data can identify an organization's true structure.

1.4 Contributions

The contributions outlined in this thesis show improvements to machine learning techniques applied to email analysis. 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 was calculated from this data. Some had been used in email analysis before, but many are new and specific to this data. Research investigating this problem in the past has focused on one of the two types of features, where this analysis

aggregates traffic- and social-based features. In general, the problem of job title classification has not been extensively studied. The leave-one-out cross-validation results in this thesis surpass the results of previous research on the Enron corpus. Finally, a closer look at email relationships lends insight into the organic structure of the organization.

1.5 Outline

This thesis continues by discussing the related works in Chapter 2. Chapter 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 Chapter 4. The results of the analysis are presented in Chapter 5. Chapter 6 presents opportunities for future work, and Chapter 7 concludes the thesis.

Chapter 2

Prior Work

This chapter investigates the previous research performed in the area of email analysis. Context is established with a discussion on the importance of workplace email communication in modern society. A description of the benchmark in email datasets, the Enron email corpus, is provided, and the drawbacks of this dataset are highlighted. Section 2.3 provides an overview of features typically used in email analysis and the works that address the problem of organization determination. A brief overview of the existing machine learning literature is presented in Section **??**.

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 [4]. 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 physical memory size and cost [5]. Out of the 600 employees involved in that study at Microsoft, the average employee had 28,660 emails stored in 133 folders; that represents a significant increase over the past ten years. Such an accumulation of emails holds a considerable amount of untapped information that could be leveraged to characterize employee roles within an organization.

2.2 Commonly Used Datasets

When the Enron email corpus was released in 2004 [6], it became the one of the largest email datasets available for public use. It is unique in that it included all of the email text, which lead to many studies involving text analysis and natural language processing. Since then, this dataset has been extensively researched on topics including spam classification [7], [8]; email categorization [9], [10]; and recipient prediction [11], [12]. However, there are known flaws and discrepancies with even the most recent versions of this dataset—ranging from misspelled email addresses [13]

to duplicate email addresses [14], and misfiled emails [15]. In one of the most popular forms of the dataset [16], 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.

There exists a list of ground truth job title labels for the Enron email corpus [17], enabling corporate hierarchy analysis. Previous work by Gilbert has made use of these labels along with performing natural language processing functions on the email text [18]. This work identified phrases more likely to be used when emailing superiors compared to phrases used when emailing peers or subordinates. The corpus and job title labels have also been used to explore patterns in office email gossip with respect to the organizational hierarchy [19]. There are clear issues with the job title labels, though. For example, Jeff Dasovic is labeled as an Employee, seemingly the lowest title. However, he had more emails than anyone else in the dataset and is identified in records as Director for State Government Affairs. There are similar inconsistencies for Sally Beck, Rick Buy, and Rod Hayslett [18]. Furthermore, out of the 161 employees with labels, 29 of them are listed as ”N/A”. Amidst all of these concerns about the Enron job title labels, this work proposes using a new dataset with accurate job titles generated from intimate knowledge of the organization.

The other datasets for analysis are rare in this vein of behavior modeling and community analysis. Other datasets used in the literature are usually either private or different types. For example, the research by Tyler et al. was performed on a internal email dataset collected by HP labs that was not publicly released [20]. Often datasets collect communication patterns from public online

sources such as Twitter [21], YouTube [22], and Wikipedia [23]. However, this research focused on organizational structure in addition to the social network components, so these datasets were unsuitable for this purpose.

2.3 Prior Email Analysis

The existing literature on analyzing social email behavior is mainly divided into two categories: traffic-based and social-based [24]. 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 about the inherent relationships. A third type of feature that has been studied utilizes email text. The work by Gilbert [18] showed that the wording used in an email could be used to make inferences about the corporate hierarchy. However, email text contains sensitive information and is often unavailable for a general email study. Therefore, features based on email text are not considered in this work.

By using features extracted from email metadata alone, previous work has been able to cluster levels of management at Enron [25]. Other metadata features such as the presence of different email attachment types and the length of emails have been shown to successfully categorize email behavior [26].

For social-based features, relational ties can be modeled as a graph 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 graphical layout [27]. A common feature used in social network analysis is betweenness centrality, which comes in several different flavors, first developed by Freeman [28]. Betweenness centrality is a measure of how many shortest paths in a graph travel over each node. In a graph with edge weights, as will be used in this work, a shortest path is a route from one node to another that has the minimum edge weight sum. Note that two or more equivalent shortest paths can exist for a pair of nodes. A node with high betweenness centrality in a social graph has been shown to represent a high degree of influence on other nodes. A betweenness centrality algorithm has been used to determine community structures within an organization [20]. There are several other types of centrality such as degree, closeness, current-flow closeness, and current-flow betweenness [29]. Avrachenkov et al. evaluated variations of betweeness centraltiy on the Enron dataset and other social networks [30]. Furthermore, Agrawal et al. classified Enron employees within superior-subordinate pairs. Using degree centrality features generated more accurate labels than natural language processing methods [31]. Other research has detected the most important email users within a corporate network without using betweenness or centrality as features [32]. Examples of alternative features include: 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 the number of nodes that can reach a node i to the average distance from those nodes to i.

Instead of considering exclusively traffic-based or social-based analytics, these feature types can be be used jointly. The only example of this approach combined features such as number of

emails, response time, cliques, and degree centrality into a “Social Score”, was used to rank Enron employees [33]. Unfortunately, this work represents preliminary research and did not publish any quantifiable results. This thesis aims to expand on this work and quantify the results when applied to a new dataset.

2.3.1 Machine Learning Overview

In this era of increasing computer memory for decreasing costs, modern businesses, researchers, medical facilities, and more are constantly collecting large volumes of data. This data may come from sources including sensors, online interactions such as social media, business transactions, and mobile devices. Email data is just one example of emerging trends in the collection and storage of large-scale information, dubbed “big data”. This term refers to datasets too large for traditional data processing techniques to handle. Machine learning has rapidly evolved to meet the demand for processing capabilities required to adequately analyze big data.

Data mining and machine learning are very closely connected. Data mining involves the extracting knowledge and patterns from large databases and selecting the best tools to perform this function. Machine learning focuses on training a machine to identify patterns and relationships from the data [34]. The work presented in this thesis performs data mining tasks, some of which make use of machine learning algorithms.

Chapter 3

Email Dataset and Feature Extraction

This chapter details all of the steps taken to transform the raw data into the high-dimensional feature space used in the learning algorithm. Information is provided on the data collected and statistics about the dataset. There is a detailed description of all features used, both traffic-based and social-based.

3.1 Data Collection

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 is available [17], however there are known issues with these labels. Due to difficulties

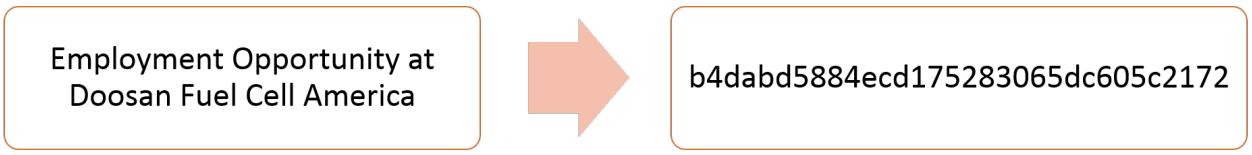

Figure 3.1: The subject on the left was hashed using the MD5 algorithm to obtain the string on the right. Hashing algorithms only function in one direction such that the original text cannot be recovered from the hashed string.

with the Enron dataset, as described in Section 2.2, this study uses a new dataset generated from emails provided by volunteers from one of the university's centers.

In consideration of the inherent privacy concerns, it was necessary to work with the Internal Review Board (IRB) to approve a data collection process which maintains participants' privacy. This dataset is meant to be representative of metadata which any company could use without divesting employees of their email privacy. 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 using the MD5 algorithm. A hashing algorithm was used to protect the email text of the participants because once text is hashed, it is impossible to recover the original message. Even if two very similar strings are hashed, their hashes will be completely independent. An example hash is shown in Figure 3.1. All email metadata was stored in a MySQL database using scripts without any researchers observing any email text. Any identifying information has been omitted from this thesis. No unauthorized persons had access to the database, which was stored on a password-protected server run by the center.

The first step in the data collection process was to design an accurate automated email parser. This script would extract the metadata from raw emails. Email formats can vary based on clients used or whether the email was sent from a mobile phone, which meant the parser script needed to be generic enough to handle many different configurations. Small issues included recognizing forwarded emails as any of the following subject prefixes: ”Fw:”, ”Fwd:”, or ”FW:”. The biggest obstacle was handling emails with embedded HTML. For each email, the parsing script first needed to recognize that there was HTML present in the body text, then accurately parse the information from the tags. Attachments were also difficult to handle because they appeared as immense blocks of encoded text. The parser was designed to identify attachments and not include that text in the character count. An example email with both HTML and an attachment is in Appendix A.

Most of the collected data is information from the header of an email such as the sending and receiving email addresses and date and time. The number of characters in the subject and body text were extracted before the subject and body text was hashed. The prefix of the email was also recorded before hashing. The hashed body and subject were loaded into the database in order to discern between emails with the same text. For example, if an email is forwarded, it will have the same hashed subject once the prefix is removed, and it is therefore possible to count the number of times an email with the same subject has been received. The number of attachments to each email was also collected. 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 whether an email was signed or encrypted was recorded. All of the metadata was collected by a

comprehensive parser script. The types of information extracted from each email are:

- 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 encrypted

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, but it involves fewer people than were used to construct the Enron dataset. While the study collected email data from only 36 volunteers, the email metadata from these volunteers identified 38 additional employees of the center. These peripheral employees were included in the study when ground truth for their job was available and when the dataset contained sufficient email metadata records (defined as at least 100 total emails). Five of the 36 direct participants were very new employees

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

	Center	Enron
Time	Nov. 2012 - Nov. 2015	Jan. 2000 - Sept. 2002
Distinct Email Addresses	32,030	75,406
Participants	36	158
Distinct Emails	579,594	252,759

to the center, and are not considered in the classification analysis due to lack of sufficient email data. The email records provided by these employees were still included to help characterize other employees. Therefore, for all analyses in Chapter 5, a total of 69 people were classified, and this set is comprised of 31 primary participants and 38 peripheral employees.

Figure 3.2 shows the email traffic of the center over time. The growth of the center between early 2013 and late 2015 is evident by the increasing baseline email traffic. Note that each year before February is a dip in email activity. This is indicative of Virginia Tech's winter break and Christmas holidays. The spikes in the graph, such as in July 2013 and February 2014, and denote when email inboxes of employees were migrated from an old mail server. For the first year of the new mail server, emails were copied over using a script which changed the timestamp of the emails to the time when they were copied. These emails were not removed from the database, however, because they represent important interactions between center employees. A more sophisticated migration process was implemented during the summer of 2014, which reduced the occurrence of such spikes.

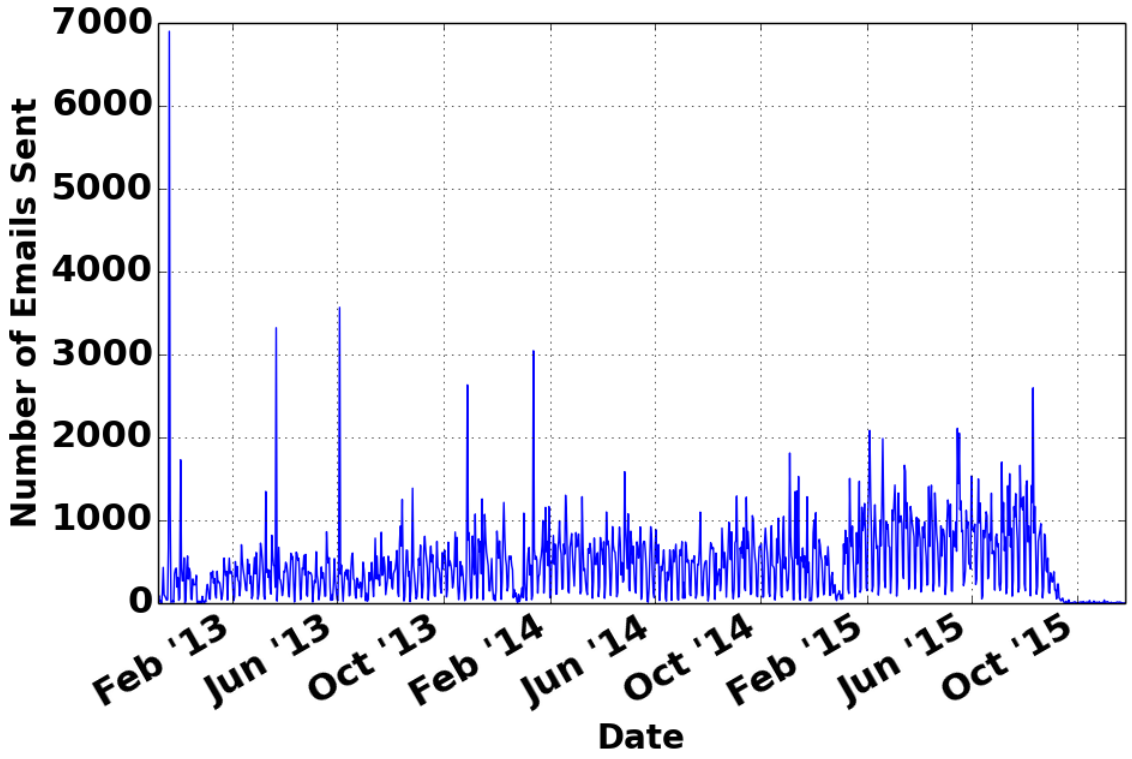

Figure 3.2: Number of emails sent in the database over time. The data trends upwards as the center has grown.

The center divides its employees into six main areas: directors, graduate students, operations, outreach, project management (PM), and research. Each person in the study was labeled with his or her job title. One important note about working with this dataset is that the distribution of these job titles is far from uniform.

A chart showing the distribution of employees into classes is shown in Figure 3.3. Approximately half of the participants are graduate students, and there are only two outreach personnel included in the study. The distributions of the study participants and peripheral employees with respect to

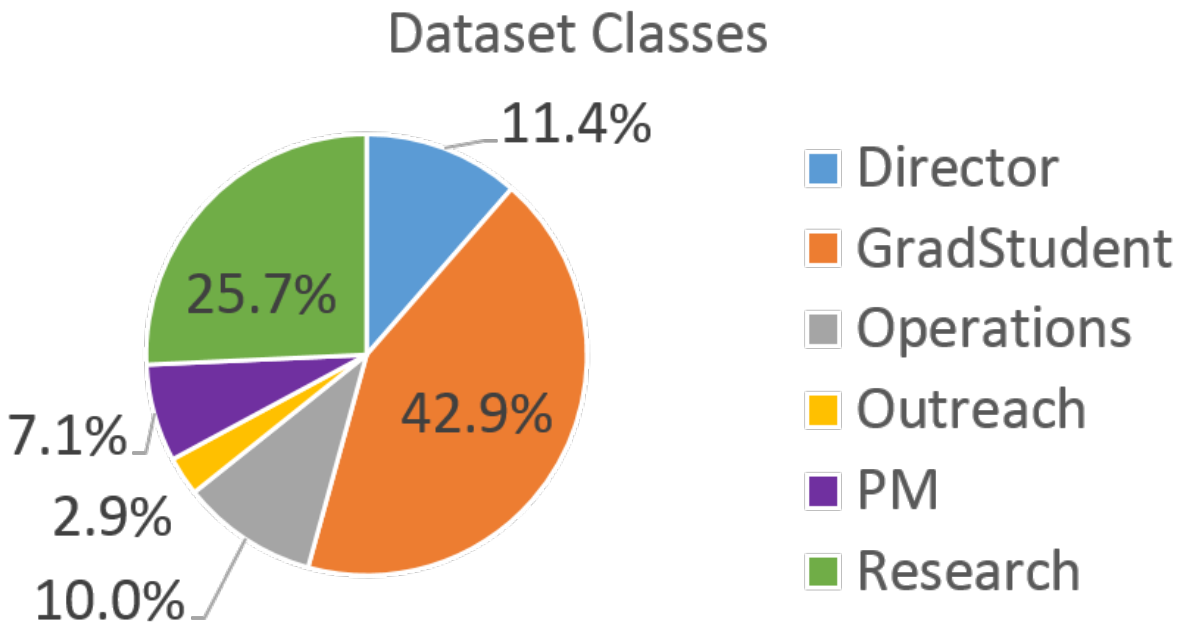

Figure 3.3: Pie chart showing the distribution of employee types in the dataset. Note that the classes are far from being uniformly distributed.

job title are very similar with one key difference. Many of the peripheral employees are former graduate students who have left the center after graduating, so the peripheral set contains two-thirds of the total graduate students. Fig. 3.4 compares the distribution of job titles to the average number of emails per person per class. Even though graduate students are by far the largest class, they send the fewest emails per person out of any class. The directors exhibit opposite behavior: there are only eight directors in the center, but collectively their emails make up almost 50% of the database.

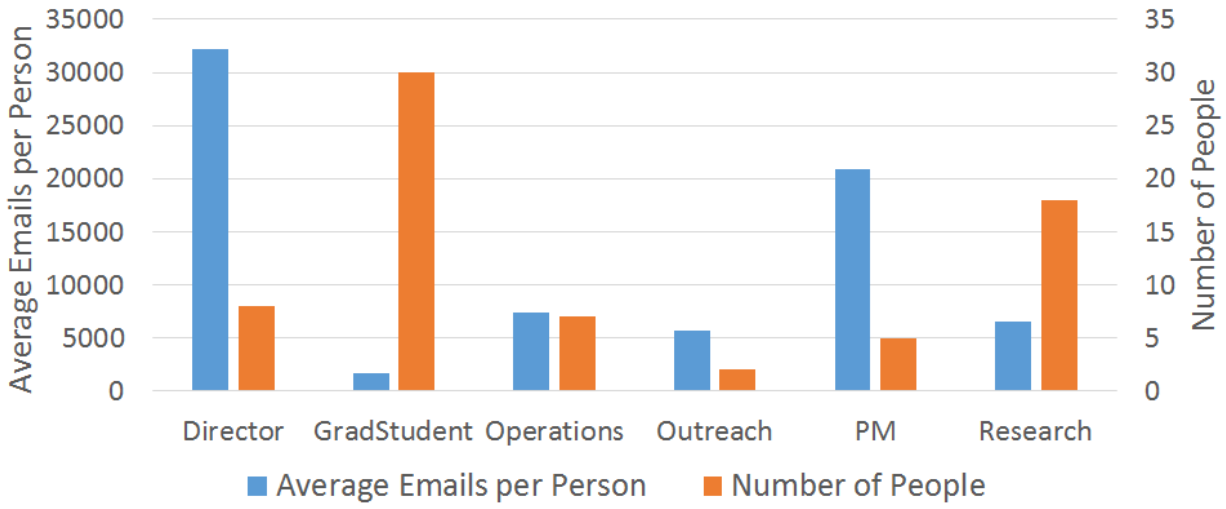

Figure 3.4: Representation of each class in the dataset with respect to average emails per person and number of people. Both distributions are very nonuniform.

3.3 Features

The study uses 114 features that were extracted from the email data: 84 traffic-based and 30 social-based. In the following sections, all features from each of the two categories are described. All of these statistics are used to characterize each employee of the center and serve as the inputs to the machine learning algorithm. A full feature list is provided in Appendix B.

3.3.1 Traffic-Based Features

The traffic-based features are those calculated purely from the collected email metadata. These metrics focus on the amount and types of emails each employee sends and receives. This section

details all of the traffic-based features used in this analysis and how they are calculated.

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 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 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 calculate 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 mutual information. This was used because it is integral to the machine learning algorithm as described in Section 4.2; details on the mutual information are provided in that section. The traffic feature with the highest mutual information was the number of forwarded emails received. Typically, those higher in the chain of command are forwarded emails in order to present them with a summary of a conversation or to pass along relevant information. Forwarding emails is also often used to pass along assignments or requests to a another employee, either up or down the chain of command [35]. Notice that this feature

also counts the raw number of emails, which means that those who send and receive more emails will in general also have a larger forwarded emails received count. It is therefore intuitive that employees who are involved in more forwarded conversations are likely to hold a higher position in the organization. Graduate students and lower-level employees are more likely to receive either replies or emails sent directly to them.

A histogram showing the different values for this metric over the different job classes is shown in Figure 3.5. The log of the raw values are plotted here for visual clarity. This figure demonstrates how difficult this classification problem can be. It is clear from the figure that graduate students received far fewer forwarded emails most any other groups. Some of the different personalities among the staff are reflected their in email behavior, which might explain why the middle bins contain employees from three or more groups. The graduate student with approximately $10^{2.5}$ forwarded emails is likley a more either more involved in office activities or at least more email-active. On the other hand, the operations employee that has less than 30 forwarded emails is clearly less email-active than other members of that team. The divisions among the classes become slightly more clear when the feature values are normalized between 0 and 1, as opposed to the raw numbers shown here. All feature values are normalized before being used by the learning algorithm.

The second best traffic feature was the number of signed emails received, and the third most important traffic feature was the number of signed emails received with unique subjects. These two values clearly will be highly correlated because they both consider signed emails, which usually signal sensitive information. Only certain groups within the center deal with this type of infor-

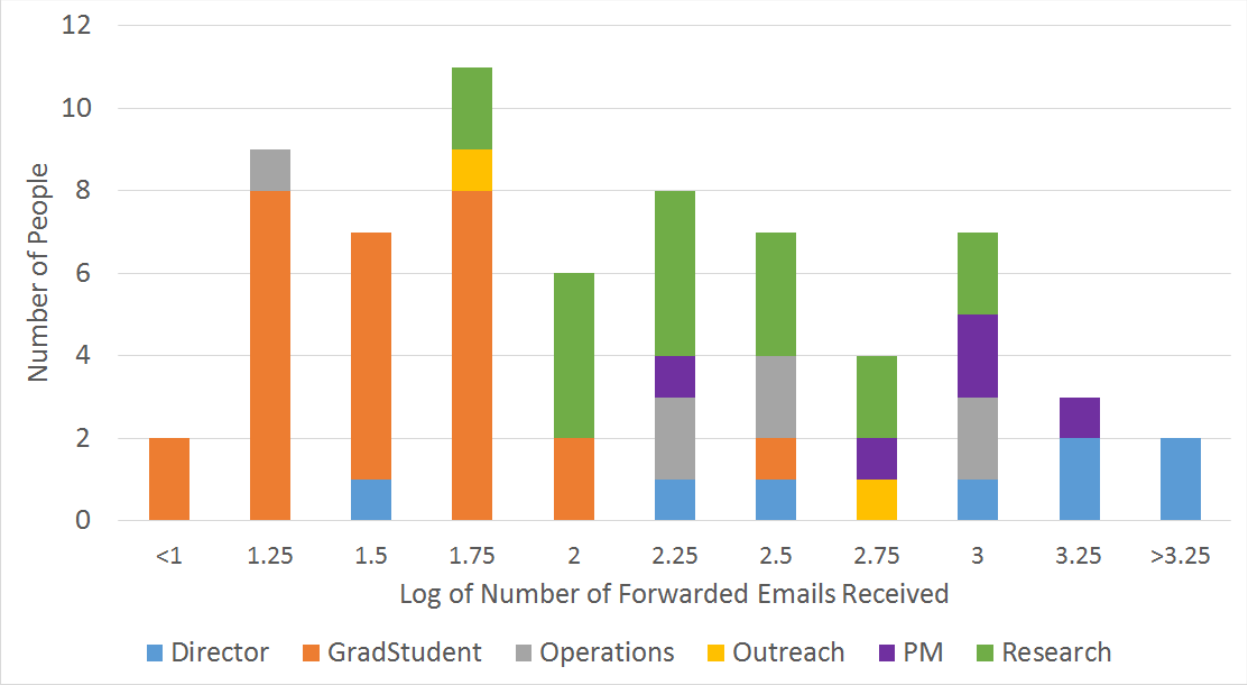

Figure 3.5: Histogram of number of forwarded emails received by job title. Note that by using different thresholds, meaningful splits in the data can be made. For example, the majority of employees with less than 100 forwarded emails are graduate students. Employees with larger values, i.e. on the order of $1,000$ forwarded emails, are likely to be higher status employees (i.e., Directors, PMs, some Research).

mation, therefore it is understandable that this feature could help divide the subjects by title. The fourth best traffic feature was the number of emails received as copies. Copying a person on a email within the workplace is often used to inform them of current activity or to inform a manager of an assignment for their subordinate [36]. Because of these behaviors, it is reasonable to anticipate directors and PMs to receive more copied emails. This is similar to the case of forwarded emails. Notice that there are intuitive explanations behind all of the features selected by the ranker.

3.3.2 Social Network Features

In addition to tracking metadata statistics, features are also derived from modeling the emails as a social network. The social features first transform email patterns in a graphical network and then calculate statistics from this model.

Social Network Representation

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 all of the center employees is shown in Figure 3.6. 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 3.7. Each of the two axes represent the employees of the center; the y-axis holds the sender of each pair and the x-axis represents the receiver. The color at each coordinate indicates how many emails

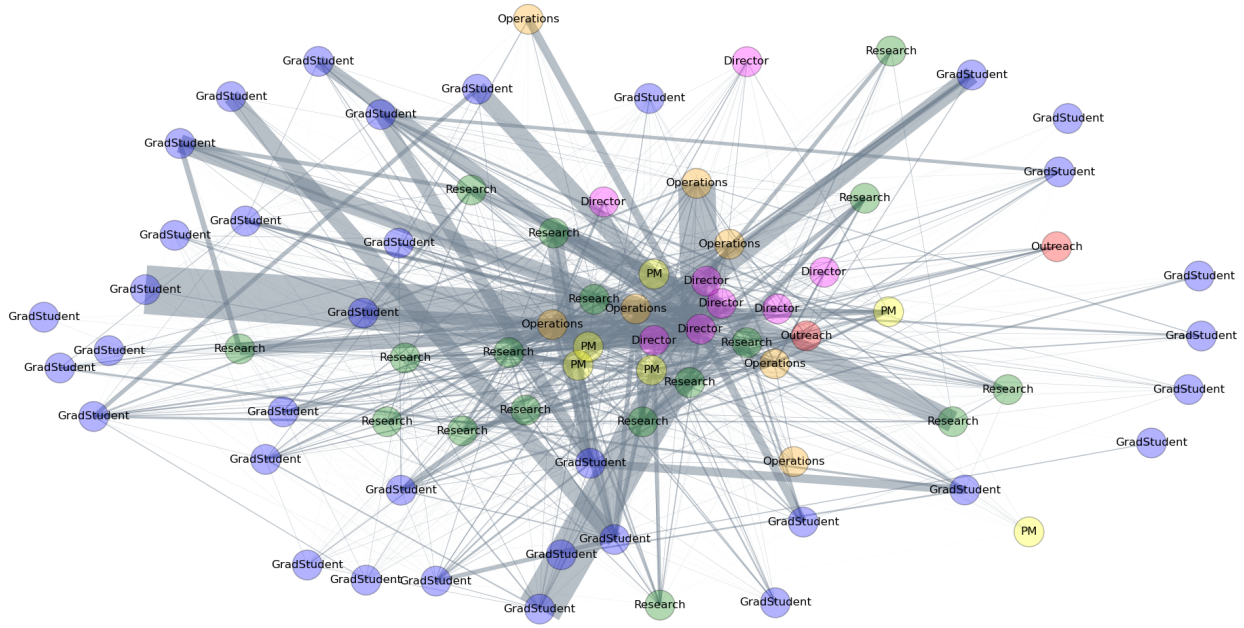

Figure 3.6: A social graph representation of the center. Nodes represent employees, and the thickness of the edges between nodes represent how many emails were exchanged.

were sent between the two employees. Some employees never exchanged any emails, while others exchanged many.

Before moving on to the features calculated from the social network model, there are several important things to notice about the adjacency matrix that will provide insight into the center's communications. First, note that Figure 3.7 is not symmetric. This means that, considering pairs of employees, one person sent more emails than the other. A metric was developed to quantify how close the adjacency matrix is to perfect symmetry, ignoring the diagonal of the matrix. This value was weighted to give high email counts more weight. For example, a pair that exchanges emails on the order of several thousands in both directions is given more weight than a pair that

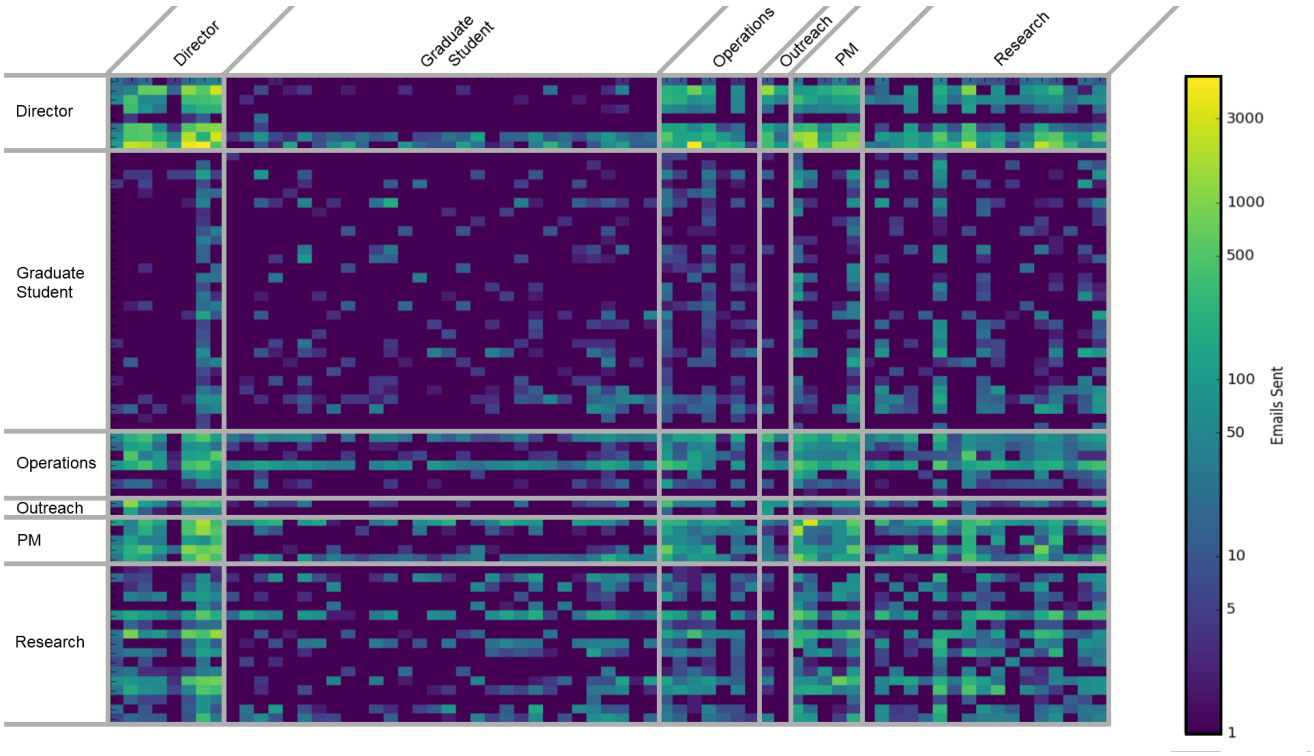

Figure 3.7: The adjacency matrix representing the social connections of the center. This graph overall is well connected and has just one component. Nonetheless, there are many pairs of individuals who never exchanged a single email.

exchanged five emails. The scale of this value is then normalized to measure symmetry such that a perfectly antisymmetric matrix would evaluate to negative one and a perfectly symmetric matrix would evaluate to one. Note that the values in the matrix represent email counts, which are always positive; therefore, the minimum value is zero. On this scale of 0 to 1, the adjacency matrix in Figure 3.7 scored 0.7036.

Another aspect of the adjacency matrix that bears mention is that the diagonal is not zero. This implies that some employees of the center email themselves. Within the dataset, 79.7% sent at least

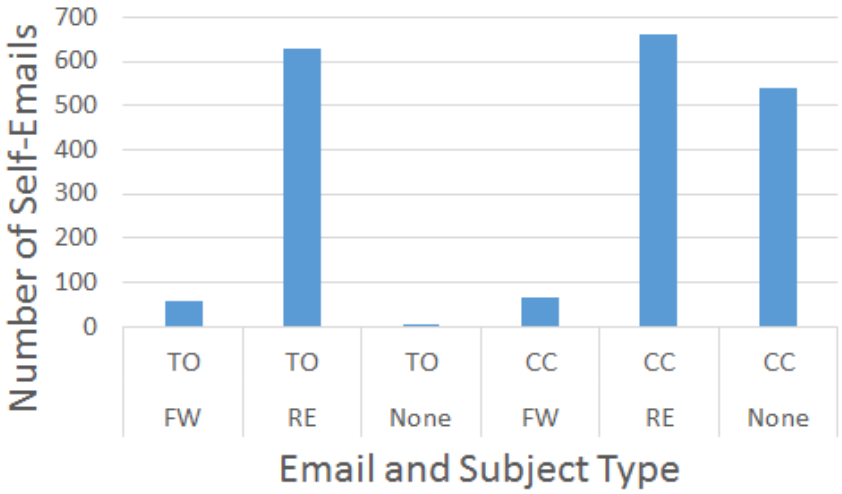

Figure 3.8: The adjacency matrix diagonal shows that some employees regularly email themselves. These emails are almost always either replies to emails or copies to themselves.

one email to themselves and 52.2% sent at least 10. These self-emails are broken down by email and subject type in Figure 3.8. Overwhelmingly these emails represent either replies to emails or copies to themselves.

Using this social network model, a suite of statistics can be calculated about the people in the graph. The graphs were modeled in python. The graph library networkx [37] was used to calculate all graph features.

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

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:

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

where $N(i)$ are the neighbors of node i and k_j is the degree of node j. The distance between nodes was 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 feature. 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) [38]. 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 [39], developed by Google 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 have been shown to predict expertise within an online social network [40].

Clustering Metrics

The triangle clustering coefficient [41] was also used as a metric. Consider a node i with neighbors m and n. The triangle clustering coefficient, C_3, 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 into this metric by finding the geometric mean [40]. 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 [42] is very similar and was also used in the algorithm. This metric, C_4, 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 $\eta_i(m,n) = 1 + q_i(m,n) + \theta_{mn}$. The indicator function θ_{mn} takes on a value of 1 if m and n are neighbors and equals 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 [43]. It is calculated as follows:

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

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. Since for this application $\mathcal{G}$ is a weighted graph, shortest path distances are calculated using Dijkstra's algorithm [44]. This is true for all of the following algorithms that consider shortest path.

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 sum of the percentage of all shortest

paths in graph $\mathcal{G}$ that traverse node i [28]. It is calculated as:

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

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 [45]:

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

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

Current flow closeness centrality, also known as information centrality, is measured for each node. In general, metrics related to current flow centrality differ from the previous centrality measures in that they consider all paths between nodes instead of exclusively shortest paths. Current flow closeness centrality in particular was modeled after how current flows in electrical networks [46]. In circuits, current is distributed over the possible paths; in this metric a similar approach 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 $\boldsymbol{B}$ 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 $\boldsymbol{B}^{-1} = \boldsymbol{C}$. From this matrix, the current flow closeness 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 $\boldsymbol{C}$, $R = \sum_{j=1}^{n} c_{ij}$. Note that because of the way it is constructed, each row in $\boldsymbol{B}$ will have the same sum. This sum is equal to the number of columns in matrix $\boldsymbol{B}$, which is equivalent to the number of nodes in graph $\mathcal{G}$, $|\mathcal{V}|$. A property of matrices states that if there exists a matrix where rows all add to the same value, λ, then all rows of the inverse of that matrix will sum to $1/\lambda$. Therefore all rows of $\boldsymbol{C}$ will sum to $1/|\mathcal{V}|$, and R is a constant that is not dependent on i [47].

Current flow betweenness centrality, as indicated by the name, also considers all possible paths between the source and target nodes. This measure is also known as random walk betweenness centrality. That is because it represents the expected number of times a random walk will cross node i when it begins at node s and ends at node t. This is calculated as:

$$b_i = \sum_{i \neq s \neq t} r_{st} \tag{3.12}$$

where r_{st} is the element of matrix $\boldsymbol{R}$ which represents the probability that a random walk from s to t passes through i.

Another feature used by the algorithm is communicability centrality, or subgraph centrality [48]. To calculate this measure, consider all of the closed walks in graph $\mathcal{G}$, of length k. Of those walks, those that begin on node i are denoted as μ_k. A closed walk is any walk that begins and ends on the

same node. Repetition of either the edges and vertices is allowed. The communicability centrality of node i, $SC(i)$ is computed as:

$$SC(i) = \sum_{k=1}^{\infty} \frac{\mu_k(i)}{k!} \tag{3.13}$$

The final social network-based feature is communicability betweenness centrality. This metric combines the concepts of current flow betweenness centrality and communicability centrality [49]. To compute the communicability betweenness centrality, consider the graph $\mathcal{G}$ with adjacency matrix $\boldsymbol{A}$. Now consider $\mathcal{G}(r)$ which is identical to $\mathcal{G}$ except with all edges connected to node r removed. The adjacency matrix of $\mathcal{G}(r)$ can be written as $\boldsymbol{A}+\boldsymbol{E}(r)$. $\boldsymbol{E}(r)$ has zeros in all elements except for those in the row and column corresponding to node r. The values in these elements are the negative weights of the removed edges, i.e. $\boldsymbol{E}(r)_{ri} = -A_{ri}$. The communicability centrality for node r is therefore:

$$\omega_r \propto \sum_p \sum_q \frac{\left(e^{\boldsymbol{A}}\right)_{pq} - \left(e^{\boldsymbol{A}+\boldsymbol{E}(r)}\right)_{pq}}{\left(e^{\boldsymbol{A}}\right)_{pq}}, p \neq q \neq r \tag{3.14}$$

The normalizing constant is omitted for clarity. Note that all metrics used in the learning algorithm were normalized to be between 0 and 1.

Most Useful Social Network Features

The feature that was ranked the most useful social-network based feature by the feature ranker in Chapter 4.3 was hubs. Recall that a node's hubs value is an estimate of how well a node is connected to authorities. These scores are calculated in an iterative reinforcement algorithm. In

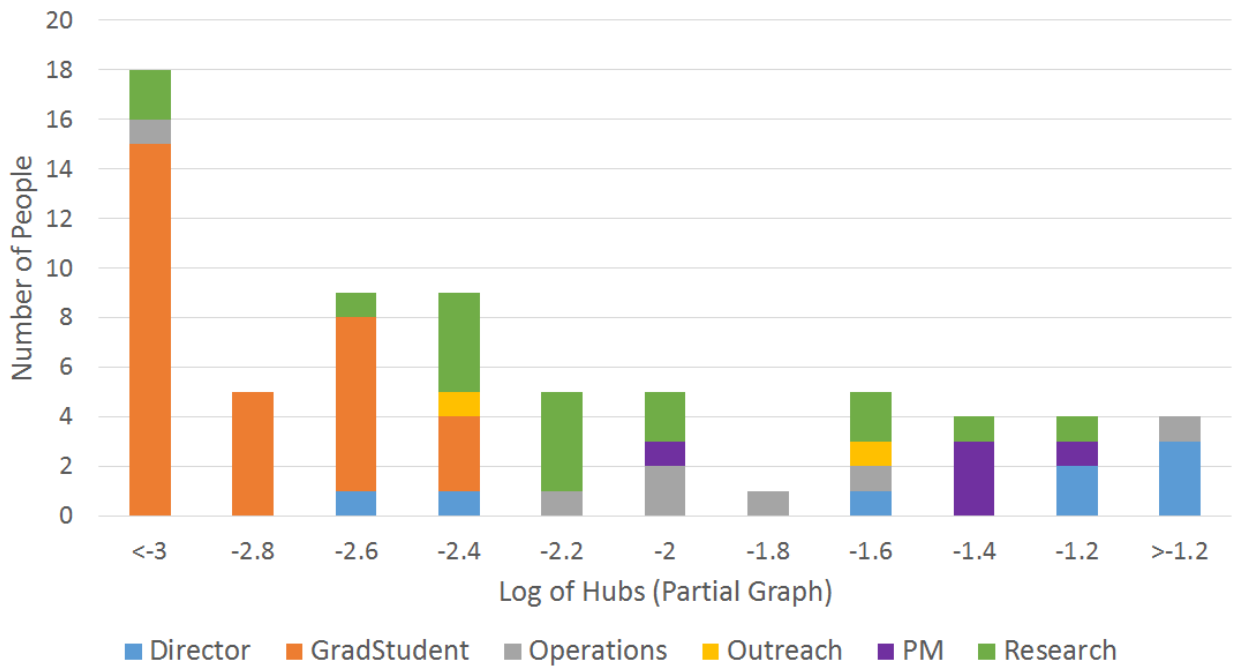

Figure 3.9: Histogram of hubs from the partial social graph by job title. Note that directors on average have the highest hub score and graduate students have the lowest. In fact, all graduate students have a full graph hubs score < 0.005. All but two PMs have a full graph hubs value > 0.025.

the original website ranking context, a hub is a site that links to many sites that are determined to be relevant or important. This can be interpreted to mean that employees with a high hubs score more often communicate with many of the directors or other important employees within the center. The histogram of hub values in the partial graph broken down by class is shown in Figure 3.9. Again, the log of the raw hubs score is presented for clarity. This figure shows more class division than Figure 3.5, as graduate students are limited to only the first four bins. It appears that the same types of employees that had high unique subjects received counts also have large full graph hubs counts, specifically most of the Directors and PMs.

The next three most useful measurements are the hubs value for the full graph, the current flow closeness centrality for the partial graph, and partial graph pagerank. These measures may seem to have less intuitive interpretations than the top-ranked traffic-based features. However, in general these webpage ranking and centrality measures attempt to capture different aspects of how well a node is embedded within the graph. In terms of this application, it measures the breadth and depth of an employee's email communications within the center.

Chapter 4

Algorithm Design

This chapter describes the analysis behind the main algorithm: the random forest. The first section describes the process of selecting the classification algorithm. The second section discusses the theory behind random forests as well as the process of training and cross-validating the model. The final section discusses the method behind ranking the different features used.

4.1 Algorithm 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. The first analysis that was performed was selecting an appropriate classification algorithm. Considering only the 32 first-ring employees,

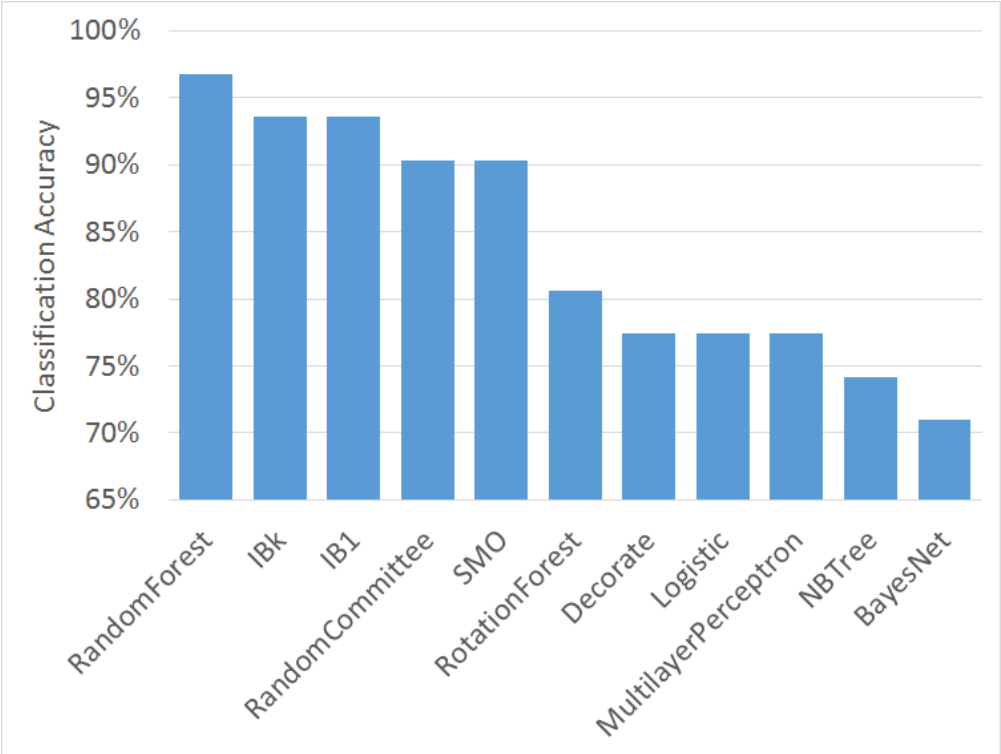

Figure 4.1: Randomly generated training and test sets were used to evaluate several different classification algorithms. Random forests were the most accurate and were therefore used for further analysis.

those with less than 100 emails were removed. The two outreach employees were also not included in this test. The emails were randomly split with equal probability into training and testing sets. Using these two sets of data, several different machine learning classification algorithms were evaluated using the java-based software package Weka [50]. The accuracy of each of these tests was recorded, and the best performers are shown in Figure 4.1. The random forest algorithm was the most accurate. This technique is popular in a variety of fields for being very robust to overfitting. Furthermore, random forests were used in the work by Namata et al. [15], which will serve as a benchmark for classification accuracy.

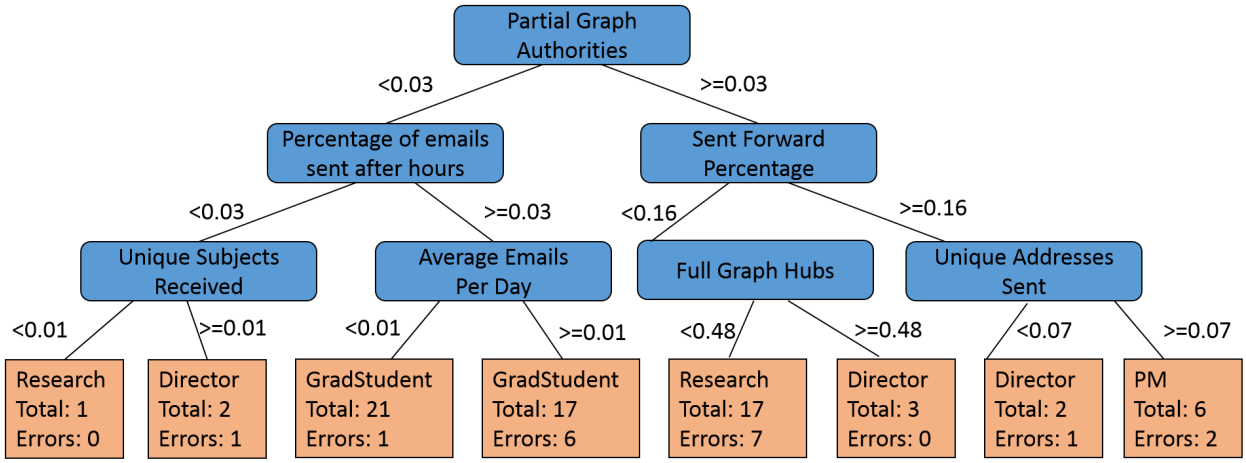

Figure 4.2: Example random tree of depth 3 to demonstrate how a few rules can be used to find significant class divisions.

4.2 Algorithm Description

Random forests training is an ensemble method of machine learning comprised of many random trees. While tree-based classifiers can be susceptible to overfitting, the random forest classifier has a lower variance and performs implicit feature selection to identify the most meaningful features. Weka uses the random forest based on the algorithm developed by Breiman [51].

A random tree is a machine learning algorithm that uses training data to divide and eventually label the data. Each rule is designed such that it will split the data in a way that maximizes the mutual information. These rules are constructed in a hierarchy that visually resembles a tree. An example random tree with depth three is shown in Figure 4.2.

Random trees are built using a greedy heuristic that determines the best split at each level. The best split is defined as the split which maximizes the mutual information and therefore makes the data after the split most homogeneous. In this model, both the class and the features are treated as random variables. The entropy of the class represents the amount of randomness in the class distribution, and is denoted as:

$$H(\text{Class}) = \sum_{l \in \text{Labels}} p(l) \log p(l) \tag{4.1}$$

The conditional entropy is evaluated for all available features. It represents the amount of randomness remaining in the class distribution when the attribute value is known. The conditional entropy is calculated as:

$$H(\text{Class}|\text{Attribute}) = \sum_{a \in \text{Attributes}} p(a) H(\text{Class}|\text{Attribute} = a) \tag{4.2}$$

$$= - \sum_{a \in \text{Attributes}} p(a) \sum_{l \in \text{Labels}} p(l|a) \log p(l|a) \tag{4.3}$$

Mutual information represents how well knowledge of the attribute informs the prediction of the class. Mutual information between the class distribution and the attribute is calculated as follows:

$$I(\text{Class}; \text{Attribute}) = H(\text{Class}) - H(\text{Class}|\text{Attribute}) \tag{4.4}$$

The random tree algorithm splits on the attribute value that maximizes the mutual information.

Note that $H(\text{Class})$ is a constant, therefore the best attribute to split on, A^*, is:

$$A^* = \arg\max_a I(\text{Class}|\text{Attribute}) \tag{4.5}$$

$$= \arg\max_a H(\text{Class}) - H(\text{Class}|\text{Attribute}) \tag{4.6}$$

$$= \arg\min_a H(\text{Class}|\text{Attribute}) \tag{4.7}$$

Note that the feature values used in this thesis are continuous, but the process above is described for discrete attribute values. The feature values can be discretized by using thresholds instead. If the feature values are sorting in ascending order, then a threshold can be defined between each consecutive pair of points. These thresholds act as discrete ways to separate the data. Therefore, the threshold that maximizes the mutual information, T^* is:

$$T^* = \arg\min_t H(\text{Class}|\text{Threshold}) \tag{4.8}$$

$$= \arg\min_t H(\text{Class}|\text{Attribute} < t\})p(\text{Attribute} < t) + H(\text{Class}|\text{Attribute} \geq t)p(\text{Attribute} \geq t) \tag{4.9}$$

A visualization of the random forest training process is shown in Fig. 4.3. Random forests build many deep random trees with imposed random variations. Individually these random trees overfit the data. To overcome this, these random trees are combined through a process of bootstrap aggregating, or bagging. The bagging process involves each random tree generating a new training set by randomly sampling people from the input training set with replacement. Each tree uses a different training set to build its rules. 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

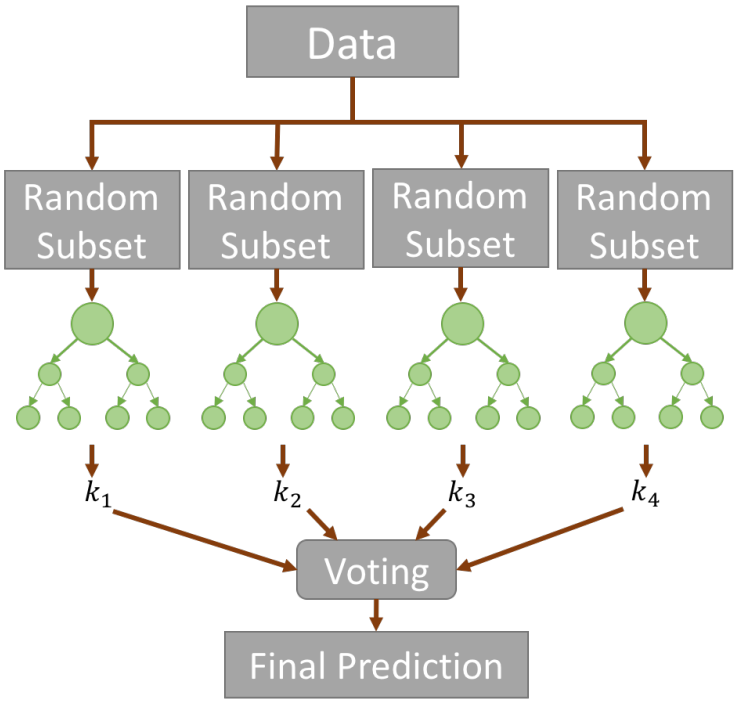

Figure 4.3: A random forest builds many random trees using subsamples of the data. Each tree generates a prediction for each test point, and the final prediction is decided by majority vote.

subsampled, so were the features. Each tree can only use a small subset of the full feature set. 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. In theory, the more meaningful features will drive some trees to make accurate predictions and will overrule trees that used the less informative features, effectively performing feature selection. Random forest training reduces the variance and increases the accuracy of the model compared to a single random tree.

Note that the way the training method is designed, the prior distributions of the classes are built into the training process. Each tree is built using a subsample of people from the dataset, and these

subsamples will reflect the original class distribution. The trees are built using these assumptions. For example, the rules might be constructed to expect more graduate students than any other category, or with the knowledge that outreach employees are relatively rare. These assumptions will be reflected in the testing data.

To use this classification algorithm, the data must be split into three groups. The emails in the dataset were randomly assigned to one of the following: 35% for training, 30% for validation, and 35% for testing. The training set was used to build the random trees. The validation set was used in lieu of a test set to optimize the parameters of the random forest. The testing set is used only once as the final evaluation of a model trained by the training set using these optimal parameters.

There are several parameters that were tuned to optimize this algorithm. The first parameter that was analyzed was the percentage splits described above. Different sectioning percentages were tried for the training and validation sets, but 35% for training and 30% for validation produced the best validation accuracy. This is likely because it is nearly an even split between the two sets. From validation analysis, the optimum number of trees in the random forest was determined to be 750. Another test found that each tree should subsample of 7 random features for maximum validation accuracy. This value represents approximately 6% of the total available features.

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 mutual information to dictate splits, mutual information was used as the evaluation criteria for the features. Each attribute was evaluated by measuring the mutual information with respect to the class, and each attribute was ranked in order of most important to least. Table 4.1 shows the top twenty features from this analysis and the features' corresponding mutual information. Highlighted rows represent the features not used previously in email analysis. Note that in this table, the features are almost perfectly split between the two types. Nine of the features are traffic-based and eleven are social-based. This highlights the importance of using both types of features in this analysis because both types of features contain valuable information related to the class of an individual.

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

Feature	Type	Information Gain
Partial graph hubs	Social	0.589
Full graph hubs	Social	0.554
Number of emails received as forwards	Traffic	0.554
Number of signed emails received	Traffic	0.514
Number of signed emails received with unique subjects	Traffic	0.514
Partial graph current flow closeness centrality	Social	0.512
Partial graph pagerank	Social	0.512
Number of emails received as copies	Traffic	0.500
Number of emails received from center employees	Traffic	0.492
Full graph current flow closeness centrality	Social	0.492
Full graph pagerank	Social	0.492
Average number of emails received per day	Traffic	0.489
Partial graph communicability centrality	Social	0.486
Partial graph communicability betweenness centrality	Social	0.486
Average number of emails per day (both sent and received)	Traffic	0.479
Number of emails sent to center employees	Traffic	0.476
Partial graph number of cliques	Social	0.470
Percentage of emails received as forwards	Traffic	0.451
Partial graph degree centrality	Social	0.448
Partial graph average shortest paths	Social	0.448

Note: Highlighted rows represent features unique to this work.

Chapter 5

Performance Analysis

The first result shows the algorithm's ability to correctly classify both the study's volunteers and the peripheral employees identified from the volunteers' emails. The second part of the results 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

As described in Section 4.2, data was split by randomly assigning each email to training or testing sets with equal probability, 0.35. The remaining 30% of the data was used as a validation set. Then, all of the metrics described in Section 3.3 were calculated for both groups separately. The training data was used as input to the random forest algorithm as described in Section 4.2, validation was performed, 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 three predictions were wrong: two people in research were misclassified as graduate students and one graduate student was misclassified as research. It is important to note that two out of the three misclassifications are peripheral employees. Therefore, the classification accuracy for the study participants is 97.3%, correctly classified peripheral employees is 93.8%, and the overall accuracy of this method using all features is 95.7% The confusion matrix for this analysis is another way to visualize the results, as shown in Figure 5.2.

On further inspection, the errors made by the learning algorithm can be explained by understanding center operations. The distinction between graduate students and research faculty is likely the least clear for two primary reasons. First, some PhD students have been at the center as long or longer than some research faculty. This leads to older PhD students having larger email counts than some of the newer research staff, which could trick the algorithm. Some research faculty are also students, which may cause their email communication patterns to more closely resemble graduate students. To further complicate matters, some graduate students joined the center as staff after they

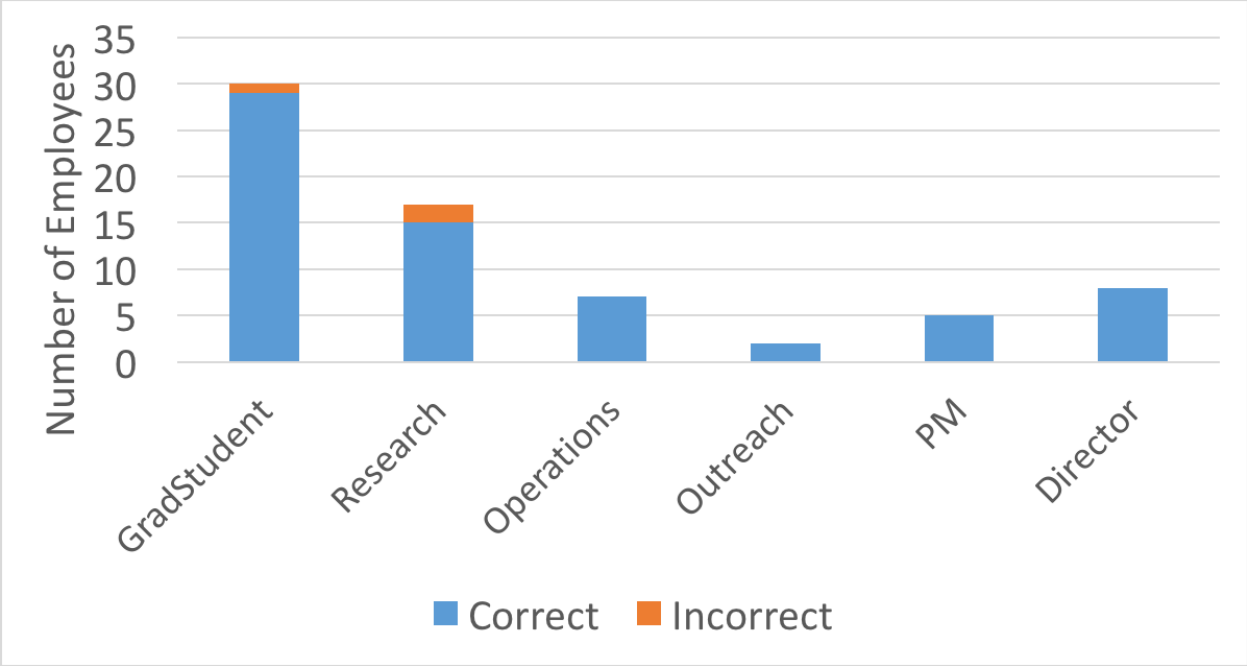

Figure 5.1: The random forest algorithm was extremely accurate even for very uneven class sizes. Note that all members of 4 classes were labeled perfectly. There were only 2 errors out of 69 employees, both of which for employees who did not provide emails for the study.

graduated.

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' email behaviors are consistent over time.

An experiment was performed to evaluate this second assumption. All employees with less than ten months of consistent email behavior were removed from this experiment. Consistent email

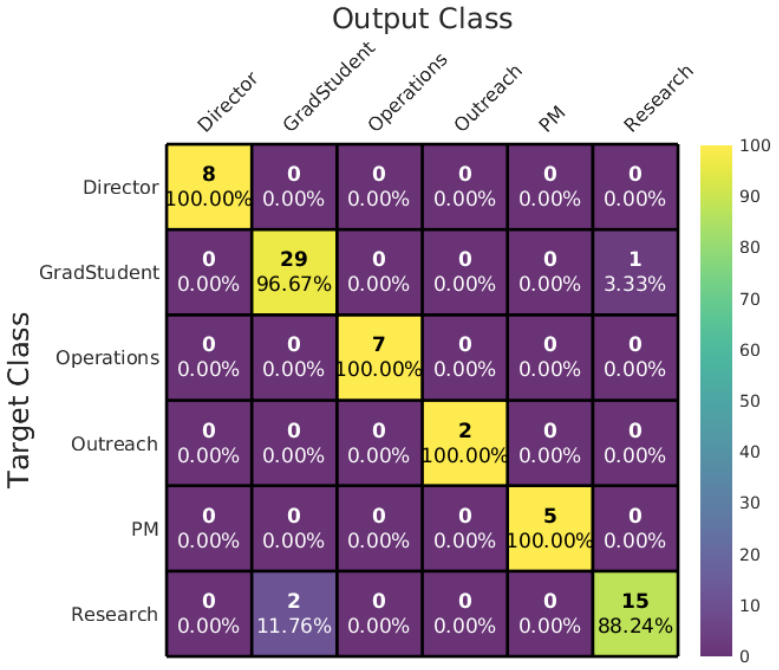

Figure 5.2: Confusion matrix for the job prediction test. The only two classes that confused the classifier were graduate student and research, making a total of 3 errors. These classes are very similar, so it is understandable that there would be errors.

behavior was defined as sending or receiving at least one email per month. One month of email data per person was set aside to act as a training set. The random forest algorithm with the parameters described in Section 4 was trained using the first month's data. Each of the nine subsequent months became a separate test set. The features were calculated for each of these months, and the accuracy of the model on the test sets was calculated. The result is a test that evaluates how the average email behavior of the center changes over time, as depicted in Figure 5.3. As the figure shows, there is a decrease in algorithm performance as the time between the training data and the test data increases, but this decrease occurs at a slow rate. Note that with six months separating the training

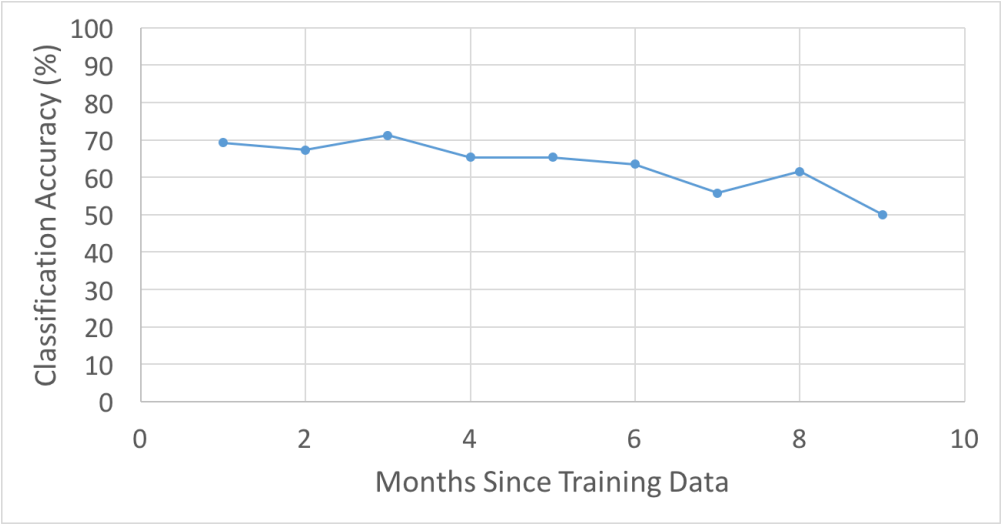

Figure 5.3: As time between the training data and the testing data increases, the accuracy of the prediction algorithm slowly decreases. This reinforces the assumption that the email behavior of the center employees is relatively constant in time.

data from the test data, the algorithm still performed with 63.4% accuracy. This is a decrease of less than 6% from when the test data is from the month immediately following the training data.

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 twenty features from Table 4.1. 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 5.4. The maximum accuracy using twenty or fewer features was achieved when either the top fifteen or sixteen were used. For this scenario, there were only 6 errors, resulting in 89.86% accuracy. This is only three more errors than was found using all 114 features. Even using just the top two features resulted in classification accuracy over 86%. Therefore, a very good classifier can

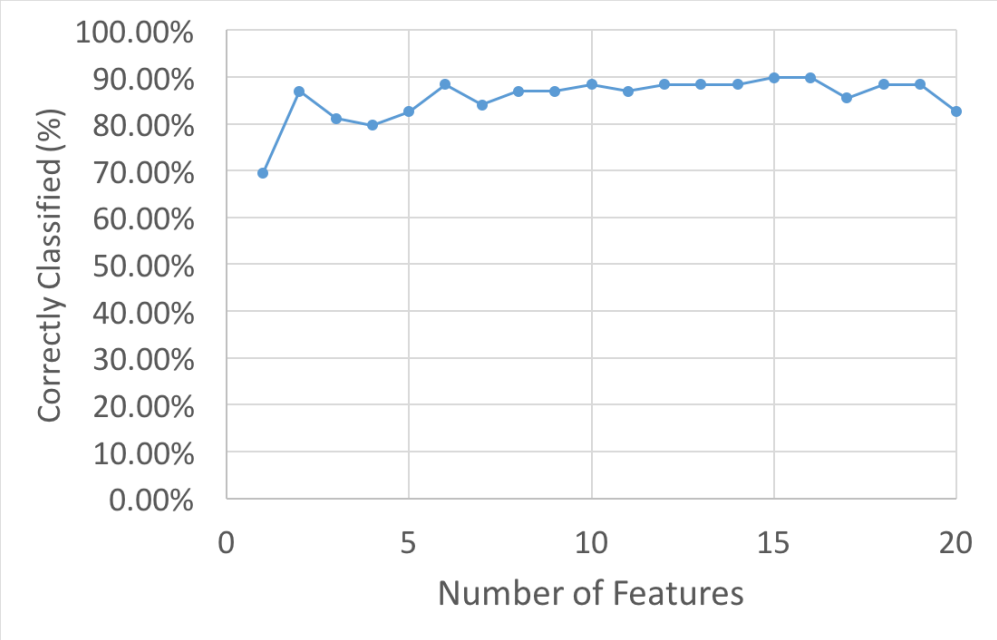

Figure 5.4: Prediction accuracy compared to number of features used for analysis. Note that the accuracy is still very high, 89.86%, when only fifteen features are used. The outcome of using only the top twenty features produces three classification errors, only one more than using the full set of 114 features.

be built using much fewer features if the features are selected properly.

5.2 Leave-One-Out Cross-Validation

Leave-one-out cross-validation (LOOCV) was used as an alternative evaluation method because it removes the time-based assumption. For LOOCV, all of the data is used to calculate the features. Then, all but one person is used as training and the left-out person is used as a single test point. For this experiment, the 2 people of the outreach class were removed because that would result in only one training point for that class. By testing each point this way, it is possible to evaluate the

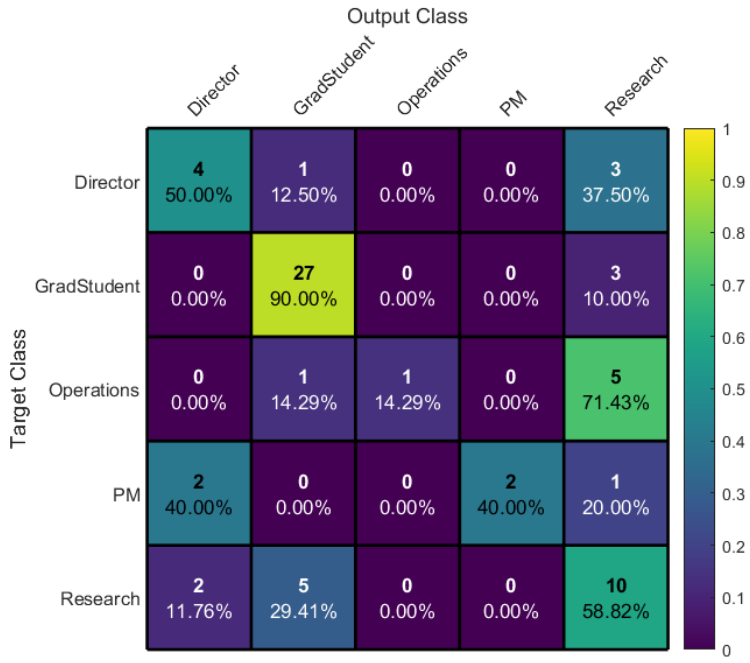

Figure 5.5: Prediction accuracy was higher for graduate students, who have more uniform behaviors and more training data. However, for each category except for research, at least half of the people classified were true members of that group.

classification algorithm without having data from all employees in both the training and test sets. This method ensures that the algorithm is not learning each person's behaviors individually. The results of this test are reported below in Figure 5.5.

This analysis resulted in a lower accuracy than the previous experiment. In total, only 65.7% were correct. Note that this performs significantly better than the naive guess approach, where the most common class is predicted for all cases. In that approach, the most common class is graduate students and that would result in 44.8% accuracy. The category of graduate student was the easiest to predict and the operations class was the most difficult. This result is understandable because

employees within operations perform the widest range of duties, including technical, administrative, and program support. One important consideration during this test is that many of the classes within the center contain employees with diverse sets of responsibilities. The director class contains directors over research, operations, and project management in addition to professors that advise students within the center. It would be plausible for a director of project management to be misclassified as a PM, or vice versa. In fact, three research directors that are professors at the university were misclassified as research staff or graduate students. There is also a wide variety of responsibilities in the research class in addition to technical work, such as mentoring graduate students and interfacing with PMs and directors. This may explain why at least one person from each other class was misclassified as research. Previous analysis showed that the boundary between graduate students and research staff can be unclear. This was clear again during this test in that four postdoctoral researchers and/or former students were misclassified as graduate students and three doctoral graduate students were misclassified as research staff. Surprisingly, when comparing the LOOCV accuracy was slightly higher for peripheral employees, 68.4% compared to 62.1% for direct participants. This is likely due to the fact that the peripheral set has more graduate students, which seems to be the easiest class to label correctly.

Namata et al. performed the same analysis on the Enron email corpus [15]. Compared to the dataset used in this analysis the Enron data involved many more people. The Enron dataset has employees distributed more evenly into six classes, only one more than used in the analysis above. Even with these advantages, the previous work was only able to achieve 62.09% LOOCV accuracy. The difference between that analysis and the work described in this thesis is that Namata et al.

only used traffic-based features. This analysis shows that integrating social-based features with the traffic-based statistics can improve job classification results.

5.3 Hierarchy Analysis

The purpose of this hierarchy analysis was to compare the email relationships from the data with the official project groups of the center. A combination of the predicted labels from Section 5.1 and the email patterns from the data generated the organic hierarchy for the center, shown in Fig. 5.6. Although the center has a very flat structure, project information often flows from directors to project managers down to research staff and finally graduate students. Outreach and operations personnel are not actively involved in project work, and are therefore omitted from this chart. The nodes of the graph represent anonymized job titles for the employees of the center. The output of the classification algorithm was used as the source of these job titles. Note that in this figure, the employees with the incorrect classification labels are outlined in red. Therefore, the two research staff are labeled as graduate students, and the mislabeled graduate student is given the title of research. Edges are drawn for each layer based on how many emails were sent to each employee in the layer above. For example, each graduate student has three edges corresponding to the three researchers they emailed most frequently. The most emailed researcher is the darkest line, the second-most common researcher is more transparent, and the third-most common link is the most transparent.

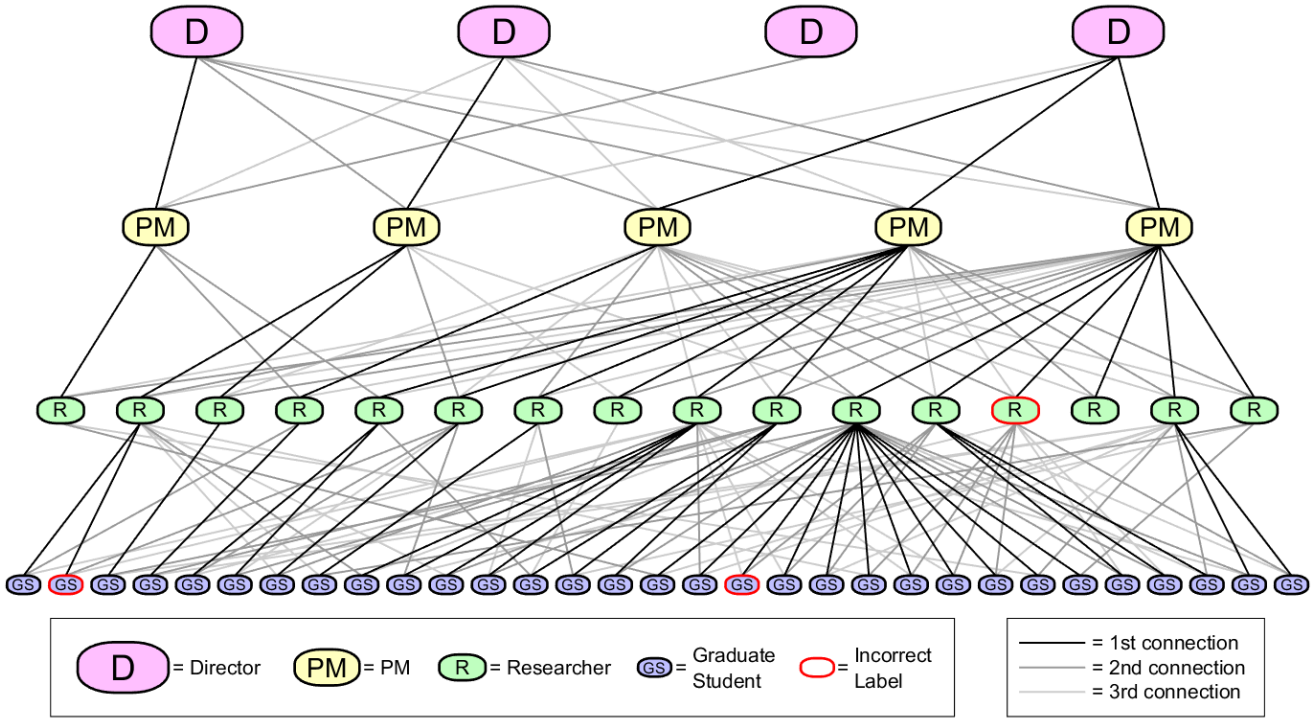

Figure 5.6: The generated organic organization chart of the center using the previous analyses. This chart could be used as a tool in corporate reorganizations to analyze workload and coworker relationships.

This type of chart could be used as a tool for understanding the underlying structure of a company in order to analyze possible reorganizations. If the center was considering dividing into two offices, this chart could help identify clusters that signal the optimal partition. Another feature of this graph is the disparity in project manager connections. Perhaps the various PMs perform different functions; some communicate more closely with researchers than others. It is also possible that some of the PMs with many researcher connections are overloaded, and the organization could benefit from redistributing projects. The final interesting thing to note about this graph is that the researcher outlined in red is in fact a graduate student. However, that person has three primary

graduate student connections. If this student were to graduate and be hired on as a research staff, this figure indicates which graduate students he or she could help mentor. This organic organization chart reveals underlying relationships within the center.

To generate a measure of how well the email-based organic structure compares to the organizational chart, the relationships in Fig. 5.6 and the rest of the email data were compared to the official hierarchy. Most of the employees at the center are organized under a director and many work with a program manager (unless for example they are a director or program manager). Therefore, from knowledge of the center, employees were labeled with their official director, and research staff and graduate students were labeled with their primary program manager. Note that these labels were not available for all participants in the study. 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. This analysis was broken down by class and plotted in Figure 5.7.

The director of each employee is predicted by the algorithm to be the director that the employee communicated with most by email. Only 52.63% 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, 91.67% of graduate students and researchers communicate

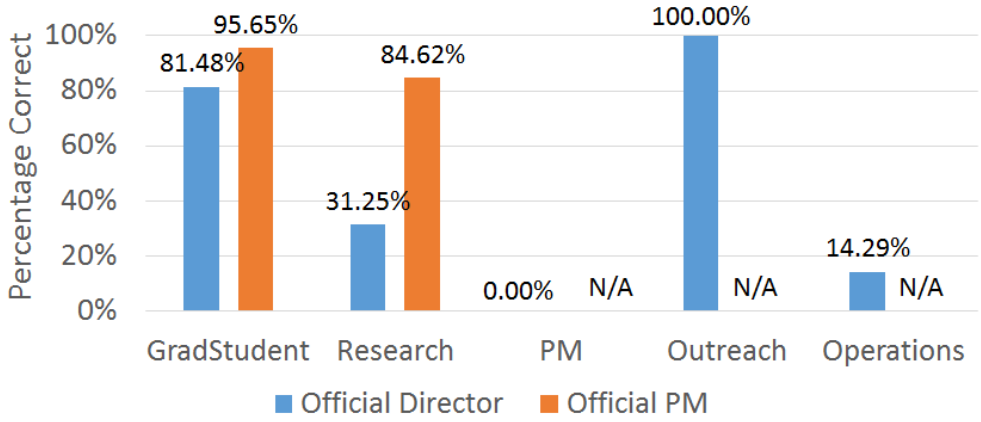

Figure 5.7: The accuracy of the director assignment appears to vary greatly between classes. However, the assigned project manager of researchers and graduate students is accurately predicted from the email data.

most frequently with their primary program manager. The relation between employees to project managers appears to be stronger than that with directors. From knowledge of the center, it appears that the few errors in this classification are due to employees who work with multiple project managers.

It appears that in general, graduate students are very likely to regularly email their assigned director and project manager. Members of the research staff communicate predictably with their primary program manager, but the director relationships from the organization chart are not reflected in the email data. Project management and operations also seem to exhibit a disconnect between the official director of each group and the directors they email.

Chapter 6

Future Work

There are obvious limitations with this dataset due to its small size. It is also not clear how well these methods and results would generalize to other email datasets, however such datasets are not publicly available. This chapter details what analyses could be possible given a larger dataset. The first section of this chapter presents some analysis showing that more data would improve the classification performance. The second section describes a deep learning method that could be used as a classifier if a massive dataset were available. The third section considers possible paths forward with this work.

6.1 Improving LOOCV Accuracy

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 5.5 shows that about one third of the predictions were incorrect. These results can typically be improved by increasing the amount of training data.

An experiment was performed to gauge how the amount of training data affects the accuracy of the LOOCV test. A series of eleven time periods were set for each person in the study, ranging from one week to ten months. Only the 37 employees who sent or received at least one email per week for ten months were included in this analysis, which is a much smaller population to test than what was used in the LOOCV analysis in Ch 5. Some of these 37 employees were direct participants in the study, but the remaining . For each employee's time slices, the features were calculated separately and then LOOCV was performed on each time slice. For example, the first run considered the most recent week of email history for each person. The LOOCV accuracy from only one week of data is obviously very low - approximately 32%. Then the test was run on the next time slice, which was the most recent two weeks. This process was repeated until LOOCV was performed on 10 months of data for each person. The results of this analysis are shown in Figure 6.1. As the number of days of data increases, the accuracy increases slowly and the variance in the accuracy decreases. A logarithmic curve fit to the data is shown. This shows that more data, as measured by time, in general increases classification accuracy.

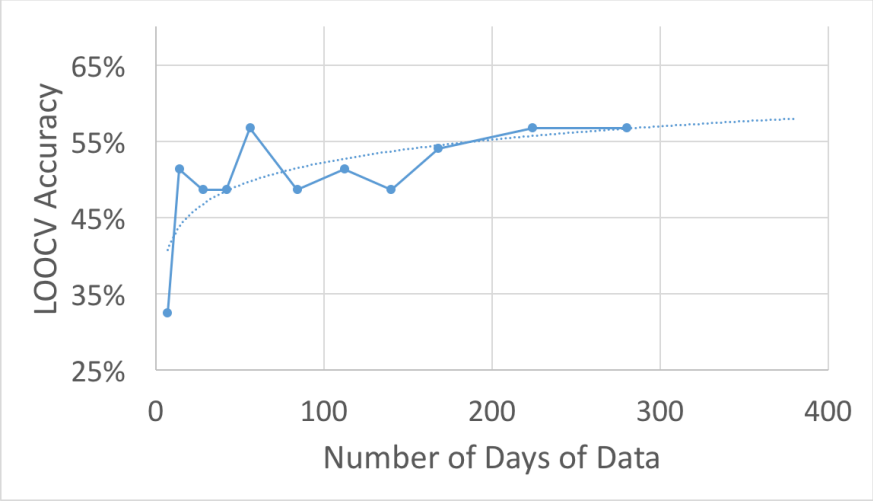

Figure 6.1: Effects of using more data for leave-one-out cross-validation. As the number of days of data increases, the accuracy increases slowly and the variance in the accuracy decreases. This implies that more email data can increase classification accuracy.

The improvement with additional data shown in Figure 6.1 is modest. Recall that the size of the population under test was decreased because of the consistent behavior constraints. In order to investigate the effect of more data, as measured in people, another test was performed. Starting with the original 67 employees (removing the two outreach employees), a random five employees were removed from training data and the LOOCV accuracy was measured for the remaining 62. To combat variance due to the randomness of removing employees, the test was repeated five times and the results averaged. Then, ten random employees were removed from the original training data and test. This continued until only seven employees remained. These results are shown in Figure 6.2. Similar trends are exhibited between this plot and the previous, namely decreasing variance and increasing accuracy. Again, a logarithmic trendline is shown for reference. It is clear that adding more people, in general, improves the performance of the classifier.

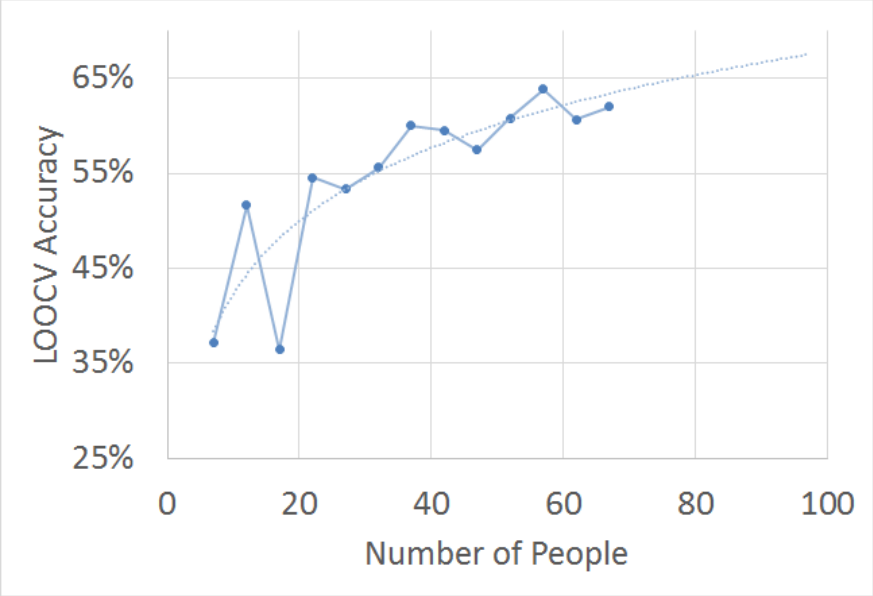

Figure 6.2: Effects of using more people for leave-one-out cross-validation. As more people are included in the training data, the LOOCV accuracy of the classifier increases slowly.

Based on these analyses, it is clear that adding more people to an email dataset and adding people with more historical email data improved the accuracy of the job title classifier.

6.2 Algorithms for Processing Larger Data

Techniques such as random forests work well for small, low-dimensional data. Machine learning is an emerging field that has proven to be very powerful and effective in solving very complex problems dealing with massive amounts of data. Unfortunately, data on that scale is unavailable for email analysis. Consider a much larger dataset, made up of millions of people without labels. In this case, a more sophisticated learning model would be necessary for successful classification

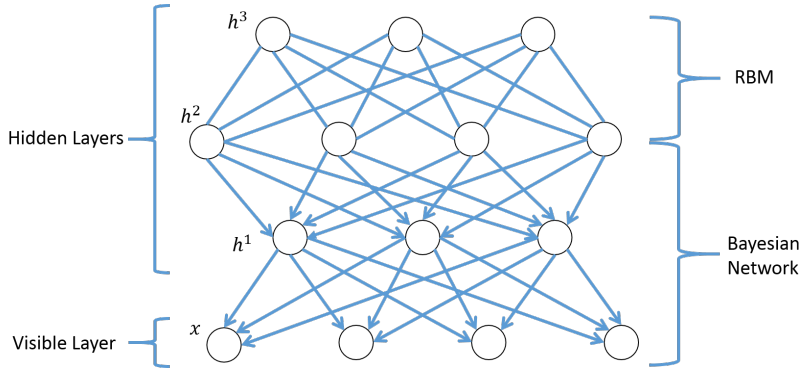

Figure 6.3: A deep belief network is a graphical network constructed by layering Restricted Boltzman machines on top of a Bayesian network [52].

and to accurately learn the complicated structure inherent to such large real-world datasets.

6.2.1 Deep Belief Networks

One powerful deep learning tools that could be used on a larger email dataset is a deep belief network. A deep belief network (DBN) is one model used to solve these types of problems. A DBN is a generative graph with nodes that represent stochastic variables. There are several layers of these nodes, some of which are visible and some of which 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. An example DBN is shown in Figure 6.3.

For the nodes in the Bayesian network layers, 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 full DBN with l hidden layers is able to model the joint distribution between inputs x and those hidden layers:

$$P(x, h^1, ..., h^l) = \underbrace{\left(\prod_{k=0}^{l-2} P(h^k|h^{k+1})\right)}_{\text{Bayesian network}} \underbrace{P(h^{l-1}, h^l)}_{\text{RBM}} \tag{6.2}$$

where $P(h^{l-1}, h^l)$ represents the joint distribution of the top two layers, which form the RBM, and $x = h_0$.

The challenge with very deep networks is that training is very difficult and very time consuming [53]. Unsupervised pre-training can improve the efficiency of training deep belief networks [54]. In that work, Hinton describes a fast, greedy method to train these networks one layer at a time. This process is as follows:

1. Train an RBM on the input layer, x, and the first hidden layer, h^1.

2. Use the RBM to transform x by sampling $p(h^1|h^0)$ or by finding the mean activations, $p(h^1 = 1|h^0)$.

3. Repeat these steps up through the network until the top layer is reached.

4. Fine-tune all parameters of the deep network

Testing on the MNIST dataset of handwritten digits, unsupervised pre-training is shown to improve classification results over only supervised training [55]. 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 [54].

In conjunction with greedy pre-training, backpropagation with some labeled data has been shown to improve both the optimization and generalization of a DBN. Suppose a representative sample of volunteers from the dataset self-identifies. This would produce a set of labeled training data, transforming an unsupervised problem into to a semi-supervised one, making backpropagation possible. Hinton et al. showed that unsupervised pre-training followed by backpropagation of a DBN outperformed traditional feed forward neural networks on the MNIST dataset [52]. The nonlinearity of the logistic function used in backpropagation mitigates the possibility of overfitting to the data.

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. Ruangkanokmas et al. developed a sophisticated version of a DBN was used to construct a sentiment classifier [56]. The purpose of that research was 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 efficiency of the algorithm. This method can classify sentiments more accurately and train more quickly than previous semi-supervised algorithms. A classifier such as this could be used to powerfully learn with extremely large email datasets, particularly those without hashed text.

6.2.2 Collective Classification

As an alternate approach to this massive data problem, statistical relational learning techniques could be used instead of deep learning. These methods are used to model, learn, and infer from relational data with complex stochastic structures [57]. Previous analysis assumed that each person behaved independently. The benefit of viewing the problem from a relational learning perspective is that information about one person may provide insight about another. However, this greatly complicates the model.

Chapter 7

Conclusions

This work presents a new dataset, approximately the size of the Enron dataset, that was collected from volunteers' emails with particular attention to protect volunteers' privacy. The new dataset includes accurate labels prepared by researchers with knowledge of the center and its employees.

A variety of features are calculated from this dataset and used with a random forest algorithm to automatically classify the center's employees. This feature set includes features that had been studied in the past but also others that are unique to this data. Furthermore, this research presents a list of features that combines the two previous schools of thought: traffic-based features and social-based features.

Random forests are shown to be powerful classifiers by predicting employee job titles with 96%

accuracy, even for employees for whom only secondhand data is available in the dataset. The result showed an improved leave-one-out cross-validation error that surpasses previous work on the Enron corpus, using a dataset that has a fraction of the samples. 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.

A method for generating an organic hierarchy was presented. This chart could be a useful tool in aiding large corporate reorganizations by clearly summarizing project-based and social relationships not immediately evident from official records. This work has shown that it is possible to glean organizational information from using carefully processed email metadata without compromising the email privacy of employees.

Chapter 8

Bibliography

[1] J. Pereira, "Northrop Grumman Realigns Sectors, Restores COO Post," *The Wall Street Journal*, Oct. 15 2015.

[2] D. B. JACOB BUNGE and C. DULANEY, "DuPont, Dow Chemical Agree to Merge, Then Break Up Into Three Companies," Dec. 11 2015.

[3] D. B. JACOB BUNGE and C. DULANEY, "Snowden Says He Took No Secret Files to Russia," Dec. 11 2015.

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

[5] 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.

[6] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *Machine learning: ECML 2004*, pp. 217–226, Springer, 2004.

[7] E. M. Bahgat, S. Rady, and W. Gad, "An e-mail filtering approach using classification techniques," in *The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt*, pp. 321–331, Springer, 2016.

[8] R. Shams and R. E. Mercer, "Classifying spam emails using text and readability features," in *2013 IEEE 13th International Conference on Data Mining (ICDM)*, pp. 657–666, IEEE, 2013.

[9] B. He, Z. Li, and N. Yang, "A novel approach for email clustering based on semantics," in *Web Information System and Application Conference (WISA), 2014 11th*, pp. 269–272, IEEE, 2014.

[10] 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.

[11] Z. Sofershtein and S. Cohen, "Predicting email recipients," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pp. 761–764, ACM, 2015.

[12] Q. Hu, S. Bao, J. Xu, W. Zhou, M. Li, and H. Huang, "Towards building effective email recipient recommendation service," in *Service Operations and Logistics, and Informatics (SOLI), 2012 IEEE International Conference on*, pp. 398–403, IEEE, 2012.

[13] A. Nordbø, "Data visualization for discovery of digital evidence in email," Master's thesis, Gjøvik University College, 2014.

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

[15] 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.

[16] 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.

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

[18] E. Gilbert, "Phrases that signal workplace hierarchy," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pp. 1037–1046, ACM, 2012.

[19] T. Mitra and E. Gilbert, "Analyzing gossip in workplace email," *ACM SIGWEB Newsletter Winter*, vol. 5, 2013.

[20] 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.

[21] R. Zafarani and H. Liu, "Social computing data repository at asu," *School of Computing, Informatics and Decision Systems Engineering, Arizona State University*, 2009.

[22] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks," in *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07)*, (San Diego, CA), October 2007.

[23] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proceedings of the 19th international conference on World wide web*, pp. 641–650, ACM, 2010.

[24] 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.

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

[26] S. Martin, B. Nelson, A. Sewani, K. Chen, and A. D. Joseph, "Analyzing behavioral features for email classification," in *CEAS*, 2005.

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

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

[29] A. Balinsky, H. Balinsky, and S. Simske, "Rapid change detection and text mining," in *Proceedings of the 2nd Conference on Mathematics in Defence (IMA), Defence Academy, UK*, 2011.

[30] K. Avrachenkov, N. Litvak, V. Medyanikov, and M. Sokol, "Alpha current flow betweenness centrality," in *Algorithms and Models for the Web Graph*, pp. 106–117, Springer, 2013.

[31] A. Agarwal, A. Omuya, A. Harnly, and O. Rambow, "A comprehensive gold standard for the enron organizational hierarchy," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pp. 161–165, Association for Computational Linguistics, 2012.

[32] G. Wilson and W. Banzhaf, "Discovery of email communication networks from the enron corpus with a genetic algorithm using social network analysis," in *IEEE Congress on Evolutionary Computation, 2009*, pp. 3256–3263, IEEE, 2009.

[33] 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.

[34] J. Fürnkranz, D. Gamberger, and N. Lavrač, *Foundations of rule learning*. Springer Science & Business Media, 2012.

[35] A. M. Bülow, J. y. H. Lee, and N. Panteli, "Distant relations: The affordances of email in interorganizational conflict," *International Journal of Business Communication*, 2016.

[36] K. Skovholt and J. Svennevig, "Email copies in workplace interaction," *Journal of Computer-Mediated Communication*, vol. 12, no. 1, pp. 42–65, 2006.

[37] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, (Pasadena, CA USA), pp. 11–15, Aug. 2008.

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

[39] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Technical Report 1999-66, Stanford InfoLab, November 1999.

[40] 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.

[41] 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.

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

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

[44] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

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

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

[47] K. Stephenson and M. Zelen, "Rethinking centrality: Methods and examples," *Social Networks*, vol. 11, no. 1, pp. 1–37, 1989.

[48] E. Estrada and J. A. Rodriguez-Velazquez, "Subgraph centrality in complex networks," *Physical Review E*, vol. 71, no. 5, 2005.

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

[50] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

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

[52] 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.

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

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

[55] 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.

[56] P. Ruangkanokmas, T. Achalakul, and K. Akkarajitsakul, "Deep belief networks with feature selection for sentiment classification," *7th International Conference on Intelligent Systems, Modelling and Simulation*, 2016.

[57] L. Getoor, *Introduction to statistical relational learning*. MIT press, 2007.

Appendix A: Example Email

```
Return-Path: <ecegrad-g+bncBDMLP5O67AFBBKOGZSXAKGQEZCOEXFI@vt.edu>
X-Original-To: kstraub@vt.edu
Delivered-To: kstraub@vt.edu
Received: from localhost (localhost [127.0.0.1])
    by hume-197.cc.ncr.vt.edu (Postfix) with ESMTP id 8BD948009D
    for <kstraub@vt.edu>; Tue, 18 Aug 2015 08:21:29 -0400 (EDT)
X-Virus-Scanned: Debian amavisd-new at hume.vt.edu
X-Spam-Flag: NO
X-Spam-Score: -2.324
X-Spam-Level:
X-Spam-Status: No, score=-2.324 tagged_above=-999 required=4
    tests=[BAYES_00=-1.9, HTML_MESSAGE=0.001, RP_MATCHES_RCVD=-0.425]
    autolearn=unavailable
Received: from hume-197.cc.ncr.vt.edu ([127.0.0.1])
    by localhost (hume-mail.ncr.vt.edu [127.0.0.1]) (amavisd-new, port 10024)
    with ESMTP id 2ojX9fo3oNED for <kstraub@vt.edu>;
    Tue, 18 Aug 2015 08:21:29 -0400 (EDT)
Received: from mr4.cc.vt.edu (mr4.cc.vt.edu [198.82.164.236])
    (using TLSv1.2 with cipher ECDHE-RSA-AES256-GCM-SHA384 (256/256 bits))
    (No client certificate requested)
    by hume-197.cc.ncr.vt.edu (Postfix) with ESMTPS id 6A69380005
    for <kstraub@vt.edu>; Tue, 18 Aug 2015 08:21:29 -0400 (EDT)
Received: from mail-qg0-f47.google.com (mail-qg0-f47.google.com [209.85.192.47])
    by mr4.cc.vt.edu (8.14.4/8.14.4) with ESMTP id t7ICLNpd000307
    for <kstraub@vt.edu>; Tue, 18 Aug 2015 08:21:28 -0400
Received: by qgdd90 with SMTP id d90sf115020074qgd.3
        for <kstraub@vt.edu>; Tue, 18 Aug 2015 05:21:23 -0700 (PDT)
X-Google-DKIM-Signature: v=1; a=rsa-sha256; c=relaxed/relaxed;
        d=1e100.net; s=20130820;
        h=x-gm-message-state:from:to:cc:subject:thread-topic:thread-index
          :date:message-id:references:in-reply-to:accept-language
          :content-language:content-type:mime-version:x-original-sender
          :x-original-authentication-results:reply-to:precedence:mailing-list
          :list-id:x-spam-checked-in-group:list-post:list-help:list-archive
          :list-unsubscribe;
        bh=Ruy4qKjNuiBq/bw588k3TUcCZhszmTlIpuPcSnzp1nY=;
        b=HbT0QllUcIJvW3Tp8NuSa94lEZ66l9jwOXutJwdyMyNz1+Na4NkU5m/WQZWPEQ+wsR
          Vk0VvSIbUAHoPof9hn+iyVa6KkxLF5lCzFNsZeU+4byNzrYkhNfGDC7wPMOkX4vxClea
          TSlzo1GxhJ48sXeClPackgsce6geijWQFOGX0J5k5RKjGJU7PiUrDTSxRRbounjn8pfy
          DBy2vw/YP83Ap6KoRGBWaIqG1NBGUxJ/FiwPNuQNzLTNbtiHFfyCjO5Xahor2LN56X8Y
          46osjoCK+mmpiy2Kyu3rr9NerAy1JjgsBdWBkUM//HpltzJ+lnHUrM7xbTER7CMsNzL6
          ZRKQ==
X-Gm-Message-State: ALoCoQl8RD1Q1iBg4zECseH0uQ/Wfu1BQxnSAcj/RX1ZQB0c0AZUPKCM/eqX7VqO9r5X3f6So+
    pnmlinUwJGa+MTj+DJd1YeBbIUqEGM5bHsWT1lTRrYkmem6YuxfykBPpgXn3rVn9k7
X-Received: by 10.140.144.17 with SMTP id 17mr5621643qhq.6.1439900458338;
        Tue, 18 Aug 2015 05:20:58 -0700 (PDT)
X-Received: by 10.140.144.17 with SMTP id 17mr5621533qhq.6.1439900458086;
        Tue, 18 Aug 2015 05:20:58 -0700 (PDT)
X-BeenThere: ecegrad-g@vt.edu
Received: by 10.140.104.201 with SMTP id a67ls962636qgf.30.gmail; Tue, 18 Aug
 2015 05:20:57 -0700 (PDT)
X-Received: by 10.140.235.129 with SMTP id g123mr12833165qhc.11.1439900457426;
        Tue, 18 Aug 2015 05:20:57 -0700 (PDT)
Received: from omr1.cc.vt.edu (omr1.cc.ipv6.vt.edu. [2607:b400:92:8300:0:c6:2117:b0e])
        by mx.google.com with ESMTPS id 62si8316831qha.12.2015.08.18.05.20.57
```

```
        for <ecegrad-g@g.mail.vt.edu>
        (version=TLSv1.2 cipher=ECDHE-RSA-AES128-GCM-SHA256 bits=128/128);
        Tue, 18 Aug 2015 05:20:57 -0700 (PDT)
Received-SPF: pass (google.com: domain of pendlelk@exchange.vt.edu designates 198.82.160.13 as
    permitted sender) client-ip=198.82.160.13;
Received: from mr3.cc.vt.edu (mr3.cc.ipv6.vt.edu [IPv6:2001:468:c80:2105:0:2b9:e1ff:8be3])
    by omr1.cc.vt.edu (8.14.4/8.14.4) with ESMTP id t7ICKvg9025677
    for <ecegrad-g@g.mail.vt.edu>; Tue, 18 Aug 2015 08:20:57 -0400
Received: from MARCONI.cc.w2k.vt.edu (marconi.cc.vt.edu [198.82.160.13] (may be forged))
    by mr3.cc.vt.edu (8.14.4/8.14.4) with ESMTP id t7ICKp3k032447;
    Tue, 18 Aug 2015 08:20:56 -0400
X-CrossPremisesHeadersFilteredBySendConnector: MARCONI.cc.w2k.vt.edu
Received: from MARCONI.cc.w2k.vt.edu (2001:468:c80:2104:0:2ce:f173:7f15) by
 MARCONI.cc.w2k.vt.edu (2001:468:c80:2104:0:2ce:f173:7f15) with Microsoft SMTP
 Server (TLS) id 15.0.995.29; Tue, 18 Aug 2015 08:20:51 -0400
Received: from MARCONI.cc.w2k.vt.edu ([fe80::a822:9a0:814d:5e06]) by
 MARCONI.cc.w2k.vt.edu ([fe80::a822:9a0:814d:5e06%12]) with mapi id
 15.00.0995.032; Tue, 18 Aug 2015 08:20:51 -0400
From: "Pendleton, Leslie" <pendlelk@exchange.vt.edu>
To: "ecegrad-g@vt.edu" <ecegrad-g@vt.edu>
CC: "Nachlas, Joel" <nachlas@vt.edu>, "Plassmann, Paul" <pep3@exchange.vt.edu>
Subject: Reliability Course
Thread-Topic: Reliability Course
Thread-Index: AQHQ2bA4ArQcMEBTPUyK5+ELpUpEgQ==
Date: Tue, 18 Aug 2015 12:20:50 +0000
Message-ID: <1439900454861.76966@exchange.vt.edu>
References: <CAG_+RUqXUSH+auz=FspLxGPcPUQ=SvdPDoGy84QkTsmuYzazQw@mail.gmail.com>,<D1F7C16F.7F323%
    pep3@vt.edu>
In-Reply-To: <D1F7C16F.7F323%pep3@vt.edu>
Accept-Language: en-US
Content-Language: en-US
X-MS-Has-Attach: yes
X-MS-TNEF-Correlator:
x-originating-ip: [128.173.88.153]
Content-Type: multipart/mixed;
    boundary="_004_1439900454861769 66exchangevtedu_"
MIME-Version: 1.0
X-OrganizationHeadersPreserved: MARCONI.cc.w2k.vt.edu
X-Gm-Spam: 0
X-Gm-Phishy: 0
X-Original-Sender: pendlelk@exchange.vt.edu
X-Original-Authentication-Results: mx.google.com;       spf=pass (google.com:
 domain of pendlelk@exchange.vt.edu designates 198.82.160.13 as permitted
 sender) smtp.mailfrom=pendlelk@exchange.vt.edu
Reply-To: pendlelk@exchange.vt.edu
Precedence: list
Mailing-list: list ecegrad-g@vt.edu; contact ecegrad-g+owners@vt.edu
List-ID: <ecegrad-g.vt.edu>
X-Spam-Checked-In-Group: ecegrad-g@vt.edu
X-Google-Group-Id: 941494946579
List-Post: <http://groups.google.com/a/vt.edu/group/ecegrad-g/post>, <mailto:ecegrad-g@vt.edu>
List-Help: <http://support.google.com/a/vt.edu/bin/topic.py?topic=25838>, <mailto:ecegrad-g+
    help@vt.edu>
List-Archive: <http://groups.google.com/a/vt.edu/group/ecegrad-g/>
List-Unsubscribe: <mailto:googlegroups-manage+941494946579+unsubscribe@googlegroups.com>,
 <http://groups.google.com/a/vt.edu/group/ecegrad-g/subscribe>

--_004_1439900454861769 66exchangevtedu_
Content-Type: multipart/alternative;
    boundary="_000_1439900454861769 66exchangevtedu_"

--_000_1439900454861769 66exchangevtedu_
Content-Type: text/plain; charset="iso-8859-1"
Content-Transfer-Encoding: quoted-printable


FYI, attached is information on a reliability theory course offered by Dr. =
Joel Nachlas in the ISE Department.

Leslie K. Pendleton, Ph.D.
Director, Student Services
Department of Electrical and Computer Engineering (Mail Code 0111)
Virginia Tech
340 Whittemore Hall
Blacksburg, VA 24061
USA
Email: pendleton@vt.edu
Phone: (540) 231-8219
```

```
Fax: (540) 231-3362

--_000_143990045486176966exchangevtedu_
Content-Type: text/html; charset="iso-8859-1"
Content-Transfer-Encoding: quoted-printable

<html>
<head>
<meta http-equiv=3D"Content-Type" content=3D"text/html; charset=3Diso-8859-=
1">
<style type=3D"text/css" style=3D"display:none"><!--P{margin-top:0;margin-b=
ottom:0;} .ms-cui-menu {background-color:#ffffff;border:1px rgb(166, 166, 1=
66) solid;font-family:'Segoe UI','Segoe WP','Segoe UI WPC',Tahoma,Arial,san=
s-serif;font-size:10pt;color:rgb(51, 51, 51);} .ms-cui-menusection-title {d=
isplay:none;} .ms-cui-ctl {vertical-align:text-top;text-decoration:none;col=
or:rgb(51, 51, 51);} .ms-cui-ctl-on {background-color:rgb(205, 230, 247);op=
acity: 0.8;} .ms-cui-img-cont-float {display:inline-block;margin-top:2px} .=
ms-cui-smenu-inner {padding-top:0px;} .ms-owa-paste-option-icon {margin: 0p=
x 6px 0px 6px;vertical-align:middle!important;padding-bottom: 2px;display:i=
nline-block;} .ms-rtePasteFlyout-option:hover {background-color:rgb(205, 23=
0, 247) !important;opacity:1 !important;} .ms-rtePasteFlyout-option {paddin=
g:8px 4px 8px 4px;outline:none;} .ms-cui-menusection {float:left; width:85p=
x;height:24px;overflow:hidden}.wf {speak:none; font-weight:normal; font-var=
iant:normal; text-transform:none; -webkit-font-smoothing:antialiased; verti=
cal-align:middle; display:inline-block;}.wf-family-owa {font-family:'o365Ic=
ons'}@font-face { font-family:'o365IconsIE8'; src:url('prem/15.0.995.34/r=
esources/styles/office365icons.ie8.eot?#iefix') format('embedded-opentype')=
,       url('prem/15.0.995.34/resources/styles/office365icons.ie8.woff') =
format('woff'),       url('prem/15.0.995.34/resources/styles/office365ico=
ns.ie8.ttf') format('truetype'); font-weight:normal; font-style:normal;}@=
font-face { font-family:'o365IconsMouse'; src:url('prem/15.0.995.34/resou=
rces/styles/office365icons.mouse.eot?#iefix') format('embedded-opentype'), =
      url('prem/15.0.995.34/resources/styles/office365icons.mouse.woff') =
format('woff'),       url('prem/15.0.995.34/resources/styles/office365ico=
ns.mouse.ttf') format('truetype'); font-weight:normal; font-style:normal;=
}.wf-family-owa {font-family:'o365IconsMouse'}.ie8 .wf-family-owa {font-fam=
ily:'o365IconsIE8'}.ie8 .wf-owa-play-large:before {content:'\e254';}.notIE8=
 .wf-owa-play-large:before {content:'\e054';}.ie8 .wf-owa-play-large {color=
:#FFFFFF/*$WFWhiteColor*/;}.notIE8 .wf-owa-play-large {border-color:#FFFFFF=
/*$WFWhiteColor*/; width:1.4em; height:1.4em; border-width:.1em; border-sty=
le:solid; border-radius:.8em; text-align:center; box-sizing:border-box; -mo=
z-box-sizing:border-box; padding:0.1em; color:#FFFFFF/*$WFWhiteColor*/;}.ie=
8 .wf-size-play-large {width:40px; height:40px; font-size:30px}.notIE8 .wf-=
size-play-large {width:40px; height:40px; font-size:30px}.notIE8 .wf-owa-tr=
iangle-down-small:before {content:'\e052';}.ie8 .wf-owa-triangle-down-small=
:before { content:'\e052';}.ie8 .wf-owa-triangle-down-small {color:#666666/=
*$WFGreyColor*/;}.wf-size-x20 {font-size: 20px!important;}--></style>
</head>
<body dir=3D"ltr" style=3D"font-size:12pt;color:#000000;background-color:#F=
FFFFF;font-family:Calibri,Arial,Helvetica,sans-serif;">
<p><br>
</p>
<p>FYI, attached is information on a reliability theory course offered by D=
r. Joel Nachlas in the ISE Department.<br>
</p>
<div>
<p><br>
</p>
<div id=3D"Signature">
<div class=3D"BodyFragment"><font size=3D"2">
<div class=3D"PlainText">Leslie K. Pendleton, Ph.D.<br>
Director, Student Services<br>
Department of Electrical and Computer Engineering (Mail Code 0111)<br>
Virginia Tech<br>
340 Whittemore Hall<br>
Blacksburg, VA 24061<br>
USA<br>
Email: pendleton@vt.edu<br>
Phone: (540) 231-8219<br>
Fax: (540) 231-3362<br>
<br>
</div>
</font></div>
</div>
</div>
<p></p>
<p><br>
```

```
</p>
<p><br>
</p>
</body>
</html>

--_000_143990045486176966exchangevtedu_--
--_004_143990045486176966exchangevtedu_
Content-Type: application/vnd.openxmlformats-officedocument.wordprocessingml.document;
    name="Fall course on reliability theory.docx"
Content-Description: Fall course on reliability theory.docx
Content-Disposition: attachment;
    filename="Fall course on reliability theory.docx"; size=13951;
    creation-date="Mon, 17 Aug 2015 20:51:41 GMT";
    modification-date="Mon, 17 Aug 2015 20:51:41 GMT"
Content-ID: <1710285A1FD5E4499D67352A0998DF58@exchange.vt.edu>
Content-Transfer-Encoding: base64

UEsDBBQABgAIAAAAIQAJJIeCgQEAAI4FAAATAAgCW0NvbnRlbnRfVHlwZXNdLnhtbCCiBAIooAAC
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAC0
lE1Pg0AQhu8m/geyVwPbejDGlPag9ahNrPG8LkPZyH5kZ/v17x1KS6qhpVq9kMAy7/vMCzOD0UqX
0QI8KmtS1k96LAIjbabMLGWv08f4lkUYhMlEaQ2kbA3IRsPLi8F07QAjqjaYsiIEd8c5ygK0wMQ6
MHSSW69FoFs/407IDzEDft3r3XBpTQAT4lBpsOHgAXIxL0M0XtHjmsRDiSy6r1+svFImnCuVFIFI
+cJk31zirUNClZt3sFAOrwiD8VaH6uSwwbbumaLxKoNoInx4Epow+NL6jGdWzjX1kByXaeG0ea4k
NPWVmvNWAiJlrsukOdFCmR3/QQ4M6xLw7ylq3RPt31QoxnkOkj52dx4a46rppLbYq+12gxAopFNM
vv6CcVfouFXuRFjC+8u/UeyJd4LkNBpT8V7CCYn/MIxGuhMi0LwD31z7Z3NsZI5Z0mRMvHVI+8P/
ou3dgqiqYxo5Bz4oaFZE24g1jrR7zu4Pqu2WQdbizTfbdPgJAAD//wMAUEsDBBQABgAIAAAAIQAe
kRq38wAAAE4CAAALAAgCX3JlbHMvLnJlbHMgogQCKKAAAgAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAjJLbSgNBDIbvBd9hyH032woi0tneSKF3
IusDhJnsAXcOzKTavr2jILpQ217m9OfLT9abg5vUO6c8Bq9hWdWg2JtgR99reG23iwdQWchbmoJn
DUfOsGlub9YvPJGUoTyMMaui4rOGQSQ+ImYzsKNchci+VLqQHEkJU4+RzBv1jKu6vsf0VwOamaba
WQ1pZ+9AtcdYN1/WDl03Gn4KZu/Yy4kVyAdhb9kuYipsScZyjWop9SwabDDPJZ2RYqwKNuBpotX1
RP9fi46FLAmhCYnP83x1nANaXg902aJ5x687HyFZLBZ9e/tDg7MvaD4BAAD//wMAUEsDBBQABgAI
AAAAIQB8O5c5IgEAALkDAAAcAAgBd29yZC9fcmVscy9kb2N1bWVudC54bWwucmVscyCiBAEooAAB
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAKyTTU+EMBCG7yb+B9K7FFZdjdmyFzXZq67x3C1T
aISWdMYP/r0VswrKogcuTWaavs/TSbtav9VV9AIejbOCpXHCIrDK5cYWgj1sb08uWYQkbS08m3Zj
FpCts+Oj1R1UksIhLE2DUUixKFhJ1FxxjqqEWmLsGrBhRztfSwqlL3gj1ZMsgC+SZMl9P4Nlg8xo
kwvmN/kpi7ZtE8h/ZzutjYJrp55rsDSC4AhE4WYYMqUvgATbd+Lgyfi4wuKAQm2Ud+g0xcrV/JP+
Qb0YXowjtRXgo6HyRmtQ1Mf/3JrySA94jIz5H6PoyL1BdPUUfjknnsILgW96V/JuTacczud00M7S
Vu6qnsdXa0ribE6JV9jd/3qVveZehA8+XPYOAAD//wMAUEsDBBQABgAIAAAAIQB7FbcDQQcAAKgo
AAARAAAAd29yZC9kb2N1bWVudC54bWzsWllz2zYQfu9M/wOGz7VF2fKliZixZSV1Gqca222mTx0I
hETEIMECkFTl13fBSzxgnUmUdvpiWSD2+vbALqhXr/8OOZpRqZiIek772HUQjYjWWTTTpOb89vTm6
dJDSOPIxFxHtOQuqnNfejz+8mnd9QaYhjTQCFpHqzmPScwKt426rpUhAQ6yOQ0akUGKsj4kIW2I8
ZoS25kL6rRO37Sb/xVIQqhTI6+NohpWTsQub3ERMI5A1FjLEWh0LOWmFWD5P4yPgHDXmHyJj64lXgOrlgybLCzgOGnRCFPcTD+XXq1zrHtrjIm84hhUeiwiQpVmbkmIWZRwWY3
aMrgQkbsE99vpZjGhTox24/bXfRc8DKJuYVm7nmSeWXT1FYMGqn7GOCYOigk3btJJCQecdBo3u4g
E5GOB8ViJPyF+YzRvAvFxn/oOa570j+9OoKky3d0jGecl160joQyVX/7Npt2xSrPhkaXc8vTy4v
24mV8VAajT8RMGmGec8hUB+pdBI7ZPPc4fUX+7MWsUN+9PQt2fGcuE7w4c5l3tPRg+ps2lk8TVWo
2HbBWOMbbsJWRoVTWrTJDFSRYbC6+u2ugvbt+Fz7Lwy+pYAQfxVqZOVaSfRzmVT7SNfRMWX7a1kT
dYvxxmRD6/Zx0h5tSwBHC4XC2Ey2tL9+rPD2cW/Om1lrrTUc24GWS91Vl+ugi+vNpNEMqtqfyYKV9
mkJJFStSshzipYTW3g1WjCCYSopDl4i0FhXkf9G2jwulKZhRZ1Q+JQfRJn7RDIUUmQmRxhsYYr1
KZyffjLBIZh30RgzPpUUQVSbAZUeRNHrCUWYEMqpXGpWLECAcTamcE4oDVPbPsVq1yD7IKKjGEPj
QbWEaDPAlb6Wg8/oCEMtDMgINgfCPwigDzTGTOZORYcLwaGkMwg7NqPoHsZUTSMMyYnMRYU/5eBN
...
```

Appendix B: Full Feature List

1. total_sent
2. unique_addresses_sent
3. unique_add_sent_perc
4. unique_subjects_sent
5. unique_sub_sent_perc
6. total_received
7. unique_addresses_received
8. uniqueadd_rec_perc
9. unique_subjects_received
10. unique_sub_rec_perc
11. total_emails
12. total_sent_signed
13. total_sent_signed_perc
14. unique_addresses_sent_signed
15. unique_add_sent_perc_signed
16. unique_subjects_sent_signed
17. unique_sub_sent_perc_signed

18. total_received_signed
19. total_rec_signed_perc
20. unique_addresses_received_signed
21. unique_add_rec_perc_signed
22. unique_subjects_received_signed
23. unique_sub_rec_perc_signed
24. total_sent_encrypted
25. total_sent_encrypted_perc
26. unique_addresses_sent_encrypted
27. unique_add_sent_perc_encrypted
28. unique_subjects_sent_encrypted
29. unique_sub_sent_perc_encrypted
30. total_received_encrypted
31. total_rec_encrypted_perc
32. unique_addresses_received_encrypted
33. unique_add_rec_perc_encrypted
34. unique_subjects_received_encrypted
35. unique_sub_rec_perc_encrypted
36. inter_hume_sent
37. inter_hume_received
38. inter_vt_sent
39. inter_vt_received
40. sent_to
41. sent_to_perc
42. sent_cc

43. sent_cc_perc
44. rec_to
45. rec_to_perc
46. rec_cc
47. rec_cc_perc
48. avg_recipients_sent
49. avg_recipients_rec
50. avg_body_chars_sent
51. avg_body_chars_rec
52. var_body_chars_sent
53. var_body_chars_rec
54. after_hours_sent
55. after_hours_sent_perc
56. after_hours_rec
57. after_hours_rec_perc
58. after_hours_sent_hume
59. after_hours_sent_hume_perc
60. after_hours_rec_hume
61. after_hours_rec_hume_perc
62. avg_sent_per_day
63. avg_rec_per_day
64. avg_emails_per_day
65. attached_sent
66. attached_rec
67. avg_attachments_sent

68. avg_attachments_rec
69. sent_re
70. sent_re_perc
71. sent_fw
72. sent_fw_perc
73. rec_re
74. rec_re_perc
75. rec_fw
76. rec_fw_perc
77. avg_subject_chars_sent
78. avg_subject_chars_rec
79. var_subject_chars_sent
80. var_subject_chars_rec
81. fg_between_centrality
82. pg_between_centrality
83. pg_avg_neighbor_degree
84. fg_avg_neighbor_degree
85. pg_clustering
86. fg_clustering
87. pg_closeness_centrality
88. fg_closeness_centrality
89. pg_degree_centrality
90. fg_degree_centrality
91. pg_current_flow_closeness_centrality
92. fg_current_flow_closeness_centrality

93. pg_current_flow_betweenness_centrality
94. fg_current_flow_betweenness_centrality
95. pg_communicability_centrality
96. fg_communicability_centrality
97. pg_communicability_betweenness_centrality
98. fg_communicability_betweenness_centrality
99. pg_load_centrality
100. fg_load_centrality
101. pg_square_clustering
102. fg_square_clustering
103. fg_eccentricity
104. pg_eccentricity
105. pg_pagerank
106. fg_pagerank
107. pg_hubs
108. pg_authorities
109. fg_hubs
110. fg_authorities
111. pg_avg_shortest_paths
112. fg_avg_shortest_paths
113. pg_num_cliques
114. fg_num_cliques