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

Kayla Straub
Virginia Tech, Blacksburg, Virginia
kstraub@vt.edu

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. For example, after the Enron dataset was made available, researchers were able to provide great insight into employee behaviors based on the available data despite the many challenges with that dataset. The work in this paper demonstrates the application of a suite of methods to an appropriately anonymized email dataset created from volunteers' email metadata. This new dataset, from an internal email server, is first used to validate machine learning and feature extraction algorithms and then to generate insight into the interactions within the center. Based solely on email data, a random forest modeled behavior patterns and accurately classified not only participants in the study but also other members of the center who were connected to participants through email. Furthermore, the data revealed relationships not present in the formal operating structure. The result is a much fuller understanding of the center's internal structure than can be found in the official organization chart.

1 Introduction

Talk about applications of determining work structure from email behavior?
Why problem is important/hard/unsolved

Organization charts do not always accurately reflect relationships within workplace. This paper seeks to determine if true behaviors are instead exhibited in corporate email communication patterns.

Email is a pervasive medium for communication in modern society - particularly in the workplace. In 2015, there were estimated over 2.6 billion email users, and 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 sends or receives 112 emails per day, accounting for 54.7% of worldwide email traffic [Radicati and Levenstein, 2015]. Retention of large email archives has become common practice with decreasing memory size and cost [Fisher *et al.*, 2006]. In that study, out of 600 employees at a high-tech company, the average employee had 28,660 emails stored in 133 folders which is a significant increase from 10 years earlier. This trove of interesting information can be leveraged to analyze the relationships between coworkers.

The following section summarizes related works in this area. Section 3 describes the process of data collection and some statistics of the dataset. The features extracted from the data are described in Section 4, and the analyses using these features are covered in Section 5. The results of the analysis are presented in Section 6. Section 7 concludes the paper and presents opportunites for future work.

2 Related Works

Email has been a common research topic over the past decade. This is in part because the Enron email dataset was released in 2004 [Klimt and Yang, 2004], allowing researchers access to a rich collection of real-world corporate emails. This dataset has been extensively researched on topics including spam classification [Martin *et al.*, 2005], [Bahgat *et al.*, 2016], [Shams and Mercer, 2013]; email categorization [He *et al.*, 2014], [Keila and Skillicorn, 2005]; and recipient prediction [Sofershtein and Cohen, 2015], [Hu *et al.*, 2012]. However, there are known flaws and discrepancies with even the most recent versions of this dataset – ranging from misspelled email addresses [Nordb, 2014] to duplicate emails [Waterman and Bruening, 2014]. In one of the most popular forms of the dataset, [Shetty and Adibi, 2004a], 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 mis-

takenly recorded as both.

The existing literature on analyzing social email behavior is mainly divided into two categories: feature-based and social-based [Tang *et al.*, 2013]. Feature-based methods calculate statistics based only on email patterns while social-based methods extract information from representing the email traffic as a social graph.

Using features extracted from email metadata, [Yelupula and Ramaswamy, 2008] was able to cluster levels of management at Enron. In addition to email traffic statistics, using features such as the presence of different email attachment types and the length of emails were shown to successfully categorize email behavior in [Martin *et al.*, 2005].

Relational ties can be modeled as a graph network where nodes represent people and edges represent interactions. This is a useful model because many statistics can be calculated from the layout of a social graph [Wasserman and Faust, 1994]. A common metric that has been shown to indicate importance in a social graph is betweenness centrality, which comes in several different flavors, and was first developed by [Freeman, 1977]. Betweenness centrality is a measure of how many shortest paths in a graph travel over each node. A node with high betweenness centrality in a social graph has been shown to represent a high degree of influence on other nodes. As [Tyler *et al.*, 2003] shows, a betweenness centrality algorithm can be used on a social graph to determine community structures within an organization. However, other metrics have been used successfully as well. For example, [Wilson and Banzhaf, 2009] detected the most important email users within a corporate network without using betweenness as a feature. Instead, they used: degree, the number of edges connected to a node; density, the ratio of actual edges to the number of possible edges; and proximity prestige, the ratio of nodes that can reach a node i to the average distance from those nodes to i.

Some research has been done in trying to combine the two different types of features. An example of this approach is seen in [Rowe *et al.*, 2007], which combined features such as number of emails, response time, cliques, and degree centrality into a "Social Score" which was used to rank Enron employees. The purpose of this paper is to further unite the two branches of research by aggregating old and novel email traffic statistics with social graph features and applying them to a new, clean dataset.

3 Data Collection

For over the past decade, the Enron dataset has been widely used to study email behaviors because it is one of the only datasets available comprised of real-world corporate emails. A list of ground truth job titles was compiled by [Shetty and Adibi, 2004b]. However, there are issues with these labels. For example, Jeff Dasovich had the most emails out of any employee in the database, and is labeled as "employee". In reality, Jeff Dasovich served as the Director for State Government Affairs. Additionally, over the period that the dataset covers, Enron was undergoing turmoil where directors changed and job titles shifted. Instead of working with uncertain labels, we decided to generate a new dataset from an organization with which we had intimate knowledge, our own center.

The center divides its employees into six main areas: directors, graduate students, operations, outreach, and research. These labels are very stable. Furthermore, we understand the intimate details for the rare case where the label has changed, i.e. when a graduate student graduates and gets hired on as a researcher. We also have inside knowledge as to personal relationships within the center which may be revealed in the data.

An Internal Review Board approved the study after reviewing a thorough proposal. Each of the 37 participants signed a waiver that detailed the specifics of the study. The following information was extracted from participants' emails:

- Destination and source email address
- Email time stamp
- Subject prefix (if any)
- Hash of subject after removing prefix
- Hash of body text
- Length of subject in characters
- Length of body text in characters
- Whether email was encrypted/signed

Table 1 compares statistics between our dataset and the Enron corpus. Note that our emails involve less people over a longer period of time. Special care was taken to protect the privacy of those involved in the study. During the collection process, all subject and body text was hashed, and all email metadata was stored in a database using scripts without any researchers observing any email details. Furthermore, any identifying information has been omitted from this report.

In an effort to increase the number of employees to analyze, statistics were also calculated for former center employees or others who had known affiliation with the center. These second-ring employees were included in the study if their job title was known and if they were involved in at least 100 emails in the database. This added 32 subjects to the study, for a total of 69 employees to analyze.

	Our Database	Enron
Time Period	11/2012-11/2015	1/2000-9/2002
Distinct Email Addresses	32,118	75,406
Participants	32	149
Distinct Emails	585,096	252,759

Table 1: Comparing our database to the Enron email corpus. Our internal database is more modern, has more emails, and covers a longer time period. However, it is comprised of less people than the Enron dataset.

4 Features

In total, 102 different features were extracted from the email data: 70 traffic-based and 32 graph-based. Below, all features will be described in both categories. An in-depth description will be presented for the top two features in each category. This ranking comes from an information gain ranker was used to determine the most important features. Section 5 goes into more detail on this topic.

4.1 Traffic-Based Features

The traffic-based features were those extracted purely from email metadata. Some of these features were very basic, such as total number of emails, total sent, and total received. Other features came from the types of emails. For example, the number of emails sent and received were counted for each of the recipient types: to, cc, and bcc. Similarly, the number of emails sent and received as replies or forwards were also used as features. The number of unique email addresses communicated with and the number of unique email subjects involving each person, both sent and received, were counted as features to be used. The average number of recipients on emails sent and received for each participant were calculated. It was hypothesized that staff members had more inter-center communications while graduate students communicated more with the university. To investigate this, the number of emails sent and received from within the center and within the university, based on the domain of the email address, were calculated.

Other useful information collected from emails included the timestamp, the character counts of the subject and body, and the presence of any attachments. From the time stamp, the time of day for each email was available. Using this information, the average number of emails sent, received, and combined were calculated. The total number of emails sent and received after hours (between 6pm and 7am EST or anytime on weekends) and the number of emails sent and received after hours within the center were also used as metrics. The mean and variance of the number of characters in the subject and body were calculated. These metrics were also broken down between emails sent and received. The raw numbers of attachments sent and received were computed as well as the average number of attachments sent and received per email. Digitally signing an email appears as an attachment, so features were calculated from the total number of signed emails sent and received, unique email addresses with signed emails, and unique subjects from signed emails.

The features above in general involved raw email counts. To normalize newer employees, features were also added where possible such that the metric was calculated as a percentage of all emails. For example, the percentage of emails that after hours out of all emails that were sent or the percentage of all received emails with unique subjects compared to all received emails.

The best traffic-based feature for predicting employee status was the number of unique subjects received. This number represents the number of distinct conversations in which an individual was involved. It is intuitive that the higher the status of a person, the more conversations they would be in. This feature is shown in the histogram in Figure 1 with normalized values. The second best traffic feature was the number of emails received that were signed. Signed emails usually signal sensitive information. Only certain groups within the center deal with this type of information, therefore it is understandable that this feature could help divide the subjects by title. Finally, the third best traffic feature was the number of emails received as forwards. Typically, those higher in the chain of command are forwarded on emails where graduate students and lower-level employees are more likely to get either replies or emails sent directly to them. Notice that there are intuitive explanations behind all of the features selected by the ranker.

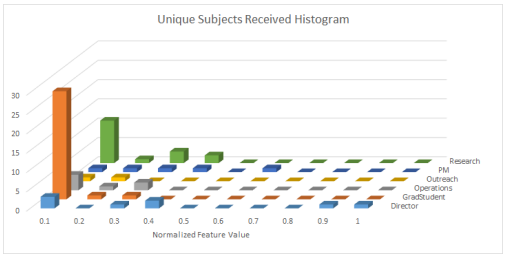

Figure 1: Histogram of unique subjects received by job title.

4.2 Social Network Features

In addition to tracking metadata statistics, features can also be derived from modeling the emails as a social network. A social network is composed of nodes, which represent people, and edges, which represent the emails between people. For this analysis, two different graphs were generated. In the full graph, an edge existed between any two individuals between which at least one email was exchanged. A second graph was generated as a subset of the first. In this partial graph, an edge was

drawn between two nodes if they had exchanged at least 10 emails. A representation of the full graph is shown below in Figure 2. This figure has two axes, both representing the employees of the center. The color at each coordinate indicates how much communication existed between the two employees. Some employees never exchanged any emails, while others exchanged many.

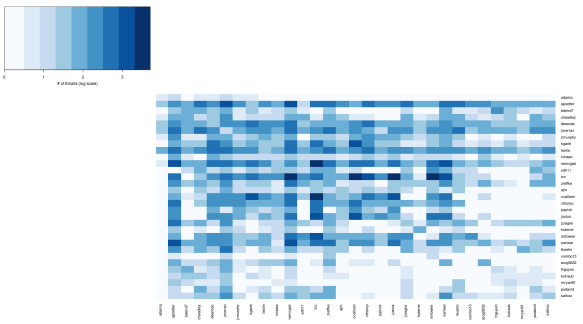

Figure 2: Adjacency matrix representing the social connections of the workplace.

With this representation, several statistics can be calculated about the people in the graph.

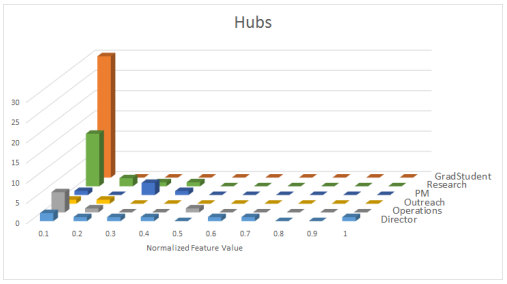

Figure 3: Histogram of hubs from social graph by job title.

5 Analysis

- Because of low sample size but high number of features, selected random forest as classification method
  - Describe Random Trees

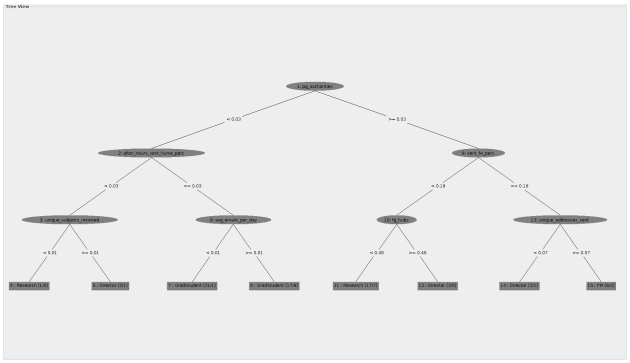

Figure 4: Example random tree of depth 3.

  - Random Forests use bootstrap AND bagging on random trees - use random components, create many, very deep trees, take a vote to classify
  - Biggest advantage: doesn't overfit
- Used information gain as feature selection method. Describe this process

Feature	Feature	Ranker
unique_subjects_received	Traffic	0.728
total_received_signed	Traffic	0.728
rec_fw	Traffic	0.719
fg_hubs	Graph	0.589
pg_communicability_centrality	Graph	0.554
pg_communicability_between_cent	Graph	0.554
rec_cc	Traffic	0.507
rec_fw_perc	Traffic	0.503
pg_degree_centrality	Graph	0.492
pg_pagerank	Graph	0.492
pg_current_flow_closeness_cent	Graph	0.492
avg_rec_per_day	Traffic	0.489
avg_emails_per_day	Traffic	0.479
pg_avg_shortest_paths	Graph	0.476
pg_closeness_centrality	Graph	0.476
unique_addresses_received_signed	Traffic	0.457
sent_cc	Traffic	0.43
rec_re	Traffic	0.43
avg_sent_per_day	Traffic	0.404

Table 2: Top 20 features ranked by the information gain method. Note that out of the to 20, there are 12 traffic-based features and 8 that are graph-based.

- Looked at prediction accuracy vs. number of features

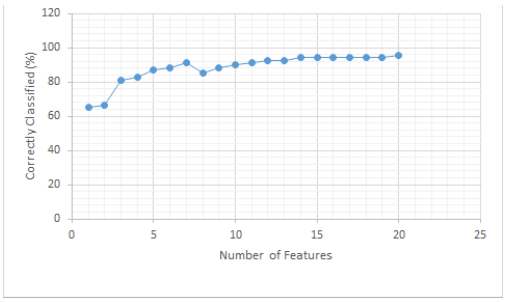

Figure 5: Prediction accuracy compared to number of features used for analysis.

- Explain splitting process (train on random % of emails, test on the rest)
- Looked into optimal train/test split:

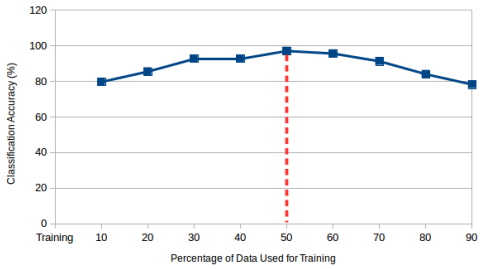

Figure 6: Prediction accuracy compared to percentage of data used for training.

- Based on the above result, I used a 50/50 split for the analysis of the classifier.

6 Results

Two tests were run on the data. In the first, the goal was to classify employees of the center based on their email behavior. In the second, the official hierarchy from the organization chart was compared to relationships displayed in the data.

6.1 Classification Results

After splitting the data randomly in half, the training data was input into the random forest algorithm described in Section 5. The model from this algorithm was used to test the remaining data and output prediction labels. These results are shown below in Figure 7. The accuracy of this method was 97.1%.

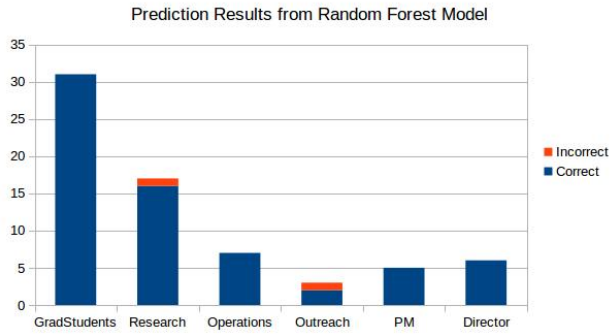

Figure 7: Prediction accuracy for test set using Random Forest model. Data was split such that 50% was used as training data and 50% was used as testing data.

It is important to note that the two misclassifications above were for second ring employees. All participants in the study were correctly classified, and 93.75% of second ring employees were correctly classified. The outreach employee is a part-time employee of the center, and was classified as a graduate student, which makes sense. There was the most training data available for graduate students, and only two training examples for outreach. Similarly, the research employee was also mislabeled as a graduate student. From our inside knowledge, we know that this person was a post-doctoral researcher. Post-docs function similarly to both researchers and graduate students, so this classification also makes sense.

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. This seems to be true as well. However, it should be noted that results did vary slightly with different random splits. Inconsistent behavior over time may have been a factor in this.

6.2 Hierarchy Analysis

The ultimate goal of this research was to determine if the relationships present in the email data confirmed or conflicted with the official organization chart. To test this, we compiled a list of the true directors for all non-director employees from the organization chart. Next, we determined which director each of these employees emailed with the most. Surprisingly, only 57.58% of employees communicate most frequently with their official director. This is understandable in cases where the director works in a different location than his or her employees. This result points to possible issues in the construction of the organization chart.

To further investigate the accuracy of the organization

chart, we also compiled a list of the project manager for graduate students and researchers in the study. In the case of an employee being on multiple projects, ground truth was selected to be the project that primarily funds the employee. Some employees in this group were not included in this list because they did not have a project manager or because they worked at the center before an official project structure was in place. This time, 72.73% of graduate students and researchers communicate most frequently with their primary program manager. The relation between employees to project managers appears to be stronger than that with directors. Many of the discrepancies can be explained by employees participating in multiple projects and therefore having multiple project managers.

7 Conclusions and Future Work

This work presented a new dataset, approximately the size of Enron, that was carefully collected from volunteers' emails with particular attention to protect participant privacy. Unlike Enron, we used our inside knowledge, to label the people in this dataset has accurate job title. A variety of statistics were calculated from this dataset, and from these statistics feature selection identified those key to making strong behavioral predictions. Random Forests were shown to be powerful classifiers for this data. They predicted employee job titles based on email data with very high accuracy, even employees that had only secondhand data was available. Our intimate knoweldge is able to interpret the few classification errors produced by the algorithm. However, there is the possibility of bias within this analysis in that training and testing are performed on the same people. Therefore, in future work we hope to futher investigate this dataset and apply this wealth of information to other problems. We also aim to evaluate these methods on other datasets, such as the Enron emails, and compare the features selected to identify any consistent predictors.

References

[Bahgat *et al.*, 2016] Eman M. Bahgat, Sherine Rady, and Walaa Gad. An E-mail Filtering Approach Using Classification Techniques. In Tarek Gaber, Aboul Ella Hassanien, Nashwa El-Bendary, and Nilanjan Dey, editors, *The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt*, volume 407, pages 321–331. Springer International Publishing, Cham, 2016.

[Fisher *et al.*, 2006] Danyel Fisher, A. J. Brush, Eric Gleave, and Marc A. Smith. Revisiting Whittaker & Sidner's email overload ten years later. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 309–312. ACM, 2006.

[Freeman, 1977] Linton C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41, 1977.

[He *et al.*, 2014] Bin He, Zefeng Li, and Nan Yang. A Novel Approach for Email Clustering Based on Semantics. In *Web Information System and Application Conference (WISA), 2014 11th*, pages 269–272, September 2014.

[Hu *et al.*, 2012] Qi Hu, S. Bao, Jingmin Xu, Wenli Zhou, Min Li, and Heyuan Huang. Towards building effective email recipient recommendation service. In *2012 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, pages 398–403, July 2012.

[Keila and Skillicorn, 2005] Parambir S. Keila and David B. Skillicorn. Structure in the Enron email dataset. *Computational & Mathematical Organization Theory*, 11(3):183–199, 2005.

[Klimt and Yang, 2004] Bryan Klimt and Yiming Yang. Introducing the Enron Corpus. In *CEAS*, 2004.

[Martin *et al.*, 2005] Steve Martin, Blaine Nelson, Anil Sewani, Karl Chen, and Anthony D. Joseph. Analyzing Behavioral Features for Email Classification. In *CEAS*, 2005.

[Nordb, 2014] Andr Nordb. Data Visualization for Discovery of Digital Evidence in Email. 2014.

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

[Rowe *et al.*, 2007] Ryan Rowe, German Creamer, Shlomo Hershkop, and Salvatore 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*, pages 109–117. ACM, 2007.

[Shams and Mercer, 2013] R. Shams and R.E. Mercer. Classifying Spam Emails Using Text and Readability Features. In *2013 IEEE 13th International Conference on Data Mining (ICDM)*, pages 657–666, December 2013.

[Shetty and Adibi, 2004a] Jitesh Shetty and Jafar Adibi. The Enron email dataset database schema and brief statistical report. *Information sciences institute technical report, University of Southern California*, 4, 2004.

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

[Sofershtein and Cohen, 2015] Zvi Sofershtein and Sara Cohen. Predicting Email Recipients. pages 761–764. ACM Press, 2015.

[Tang *et al.*, 2013] Guanting Tang, Jian Pei, and Wo-Shun Luk. Email mining: tasks, common techniques, and tools. *Knowledge and Information Systems*, 41(1):1–31, June 2013.

[Tyler *et al.*, 2003] Joshua R. Tyler, Dennis M. Wilkinson, and Bernardo A. Huberman. Email as Spectroscopy: Automated Discovery of Community Structure within Organizations. *arXiv:cond-mat/0303264*, March 2003. arXiv: cond-mat/0303264.

[Wasserman and Faust, 1994] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, November 1994.

[Waterman and Bruening, 2014] K. Krasnow Waterman and Paula J. Bruening. Big Data analytics: risks and responsibilities. *International Data Privacy Law*, 4(2):89–95, May 2014.

[Wilson and Banzhaf, 2009] Garnett Wilson and Wolfgang Banzhaf. Discovery of email communication networks from the Enron corpus with a genetic algorithm using social network analysis. In *Evolutionary Computation, 2009. CEC'09. IEEE Congress on*, pages 3256–3263. IEEE, 2009.

[Yelupula and Ramaswamy, 2008] K. Yelupula and Srini Ramaswamy. Social network analysis for email classification. In *Proceedings of the 46th Annual Southeast Regional Conference on XX*, pages 469–474. ACM, 2008.