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

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, 2004], 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.

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.

- Maybe add paragraph about algorithm-specific learning?

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 [ref labels here]. 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 were shifting. Instead of working with uncertain labels, we decided to generate a new dataset from an organization with which we had intimate knowledge.

- Motivation for new data set
  - No ground truth hierarchy for Enron
  - Cite Dasovich listed as an "Employee" when his real title was Director for State Government Affairs
  - Intimate knowledge of lab hierarchy/structure and personal interactions
- Anonymization process
- First (37) vs. second ring (32) subjects for a total of 69 employees analyzed
  - The first ring consisted of participants who signed release forms to collect metadata
  - The second ring consisted of employees who did not sign a release form, but had known affiliation with the lab. We also ensured that there were at least 100 emails in the database for each of these employees.
- Statistics
  - Data collected
    - Destination and source email address
    - Email timestamp
    - 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
  - Number of people (32118 distinct email addresses)
  - Number of emails (2,276,770 in recipients table)
  - Time period (1114 days/3.05 years, from 11/6/12-11/25/15(ish))

4 Features

In total, 102 different features were extracted from the email data: 70 traffic-based and 32 graph-based.

- Briefly talk about ranker

4.1 Traffic-Based Features

- Highlight top 3 traffic-based features

4.2 Social Network Features

- Talk about how social graphs were created

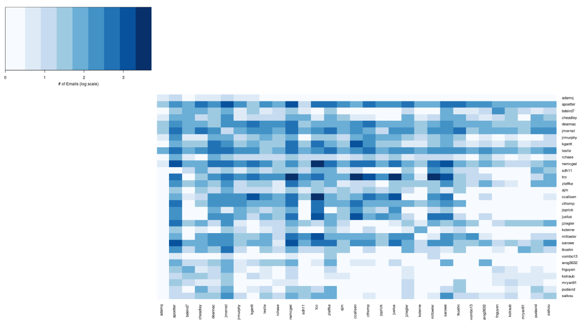

Figure 1: Adjacency matrix representing the social connections of the workplace.

- Highlight top 3 social-based features

5 Analysis

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

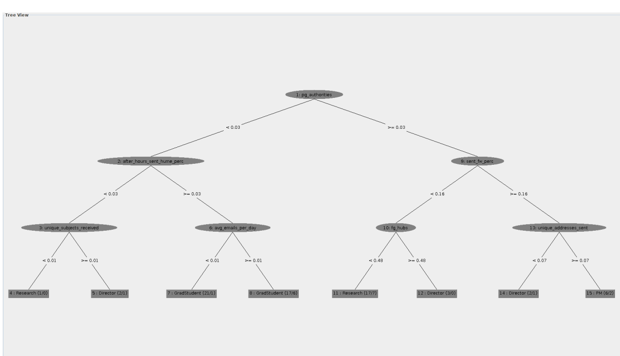

Figure 2: Example random tree of depth 3.

  - 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_colseness_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 1: 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.

- Manually looked at feature breakdown compared to classes

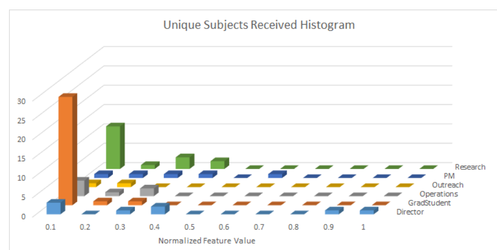

Figure 3: Histogram of unique subjects received by job title.

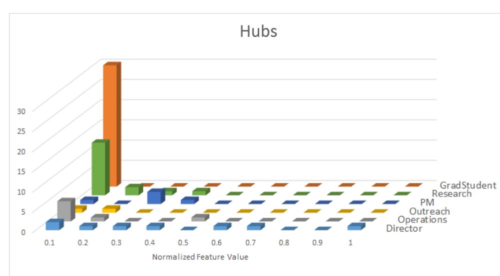

Figure 4: Histogram of hubs from social graph by job title.

- Looked at prediction accuracy vs. number of features

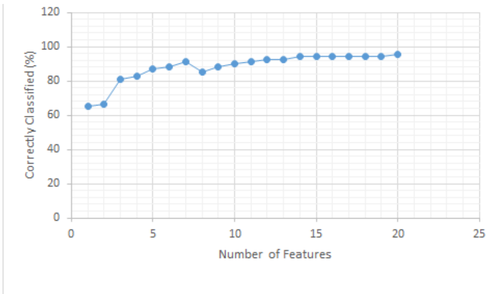

Figure 5: Prediction accuracy compared to number of features used for analysis.

6 Results

6.1 Classification Results

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

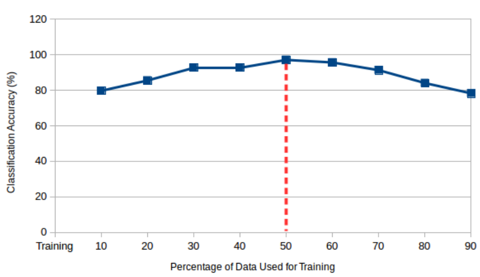

Figure 6: Prediction accuracy compared to percentage of data used for training.

- Based on the above result, I used a 50/50 split for the analysis of the classifier.
- Classification Results
- Assumption: employees with the same title exhibit
- Assumption: Email behavior is consistent over time

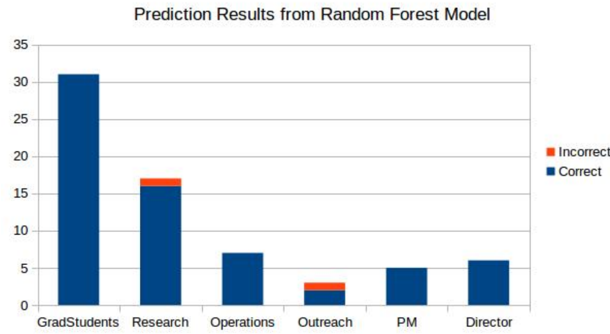

Figure 7: Prediction accuracy for test set using Random Forest model. Data was split such that 50% was used as training data and 50% was used as testing data.

- Even the errors are explainable (not much data on Outreach. The researcher that was misclassified was mislabelled as a graduate student, and this person is a post-doc.
- Both misclassifications were second-ring

6.2 Hierarchy Analysis

- Only 57.58% of employees communicate most frequently with their director from the organization chart.
- 72.73% of graduate students and researchers communicate most frequently with their primary program manager.

- Many of the discrepancies make sense to those of us with insider knowledge and reflect formal choices made in generating the org chart

7 Conclusions and Future Work

- Presented new dataset that was carefully cleaned with inside knowledge and has accurate labels.
- Random Forests are shown to be powerful classifiers for this data. They classified with high accuracy
- Used both traffic-based and graph features helped improve accuracy
- Acknowledge we are testing back on ourselves, randomness in email splitting process
- From our intimate knowledge of the workings of the lab, we know that hierarchy discrepancies can be explained by multiple projects, directors/PMs working in different offices, etc.
- Do more things with dataset
- Evaluate methods on Enron emails and compare which features are 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, 2004] 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.

[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.