Exploring Workplace Hierarchy Through Corporate Email Metadata

Kayla Straub

***Abstract*—**

I. Introduction

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 [1]]. According to [2], retention of large email archives has become common practice with decreasing memory size and cost. 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.

II. Related Works

Email has been a common research topic over the past decade. Most notably, the Enron email database was released in 2004 [3], following the company's collapse. This dataset has been extensively researched on topics including spam classification [4], [5], [6]; email categorization [7], [8]; and recipient prediction [9], [10]. However, there are known flaws and discrepancies with even the most recent versions of this dataset – ranging from misspelled email addresses [11] to duplicate emails [12]. In one of the most popular forms of the dataset, [13], the database includes 253,735 emails sent as 'CC' and 253,713 emails sent as 'BCC'. Further inspection reveals that emails sent as one type or the other were mistakenly recorded as both.

The existing literature on analyzing social email behavior is mainly divided into two categories: feature-based and social-based [14].

Using features extracted from email metadata, [15] was able to cluster levels of management at Enron. In addition to email traffic statistics, using features such as the presence of different email attachment types and the length of emails were shown to successfully categorize email behavior in [4].

Relational ties can be modeled as a graph network where nodes represent people and edges represent interactions. This is a useful model because many useful statistics can be calculated from the layout of a social graph [16]. A common metric that has been shown to indicate importance in a social graph is betweenness, which comes in several different flavors, and was first developed by [17]. As [18] shows, a betweenness centrality algorithm can be used on a social graph to determine community structures within an organization. However, other features have been used successfully as well. Degree, density, and proximity prestige were used in [19] to detect the most important email users within a corporate network.

Some research has been done in trying to combine the two different types of features. An example of this approach is seen in [20], 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 an original, clean dataset.

[Maybe add paragraph about algorithm-specific learning?]

III. Features

A. Traffic-Based Features

B. Social Network Features

IV. Analysis

[Image: Feature histograms]
[Image: Random tree rules]

V. Results

[Image: Graphical representation of the predictions]

VI. Conclusions and Future Work

References

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

[2] D. Fisher, A. J. Brush, E. Gleave, and M. A. Smith, "Revisiting Whittaker & Sidner's email overload ten years later," in *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. ACM, 2006, pp. 309–312. [Online]. Available: http://dl.acm.org/citation.cfm?id=1180922

[3] B. Klimt and Y. Yang, "Introducing the Enron Corpus." in *CEAS*, 2004. [Online]. Available: http://bklimt.com/papers/2004_klimt_ceas.pdf

[4] S. Martin, B. Nelson, A. Sewani, K. Chen, and A. D. Joseph, "Analyzing Behavioral Features for Email Classification." in *CEAS*, 2005. [Online]. Available: http://blaine-nelson.com/research/pubs/Martin-Sewani-CEAS-2005.pdf

[5] 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*, T. Gaber, A. E. Hassanien, N. El-Bendary, and N. Dey, Eds. Cham: Springer International Publishing, 2016, vol. 407, pp. 321–331. [Online]. Available: http://link.springer.com/10.1007/978-3-319-26690-9_29

[6] R. Shams and R. Mercer, "Classifying Spam Emails Using Text and Readability Features," in *2013 IEEE 13th International Conference on Data Mining (ICDM)*, Dec. 2013, pp. 657–666.

[7] 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*, Sep. 2014, pp. 269–272.

[8] 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. [Online]. Available: http://link.springer.com/article/10.1007/s10588-005-5379-y

[9] Z. Sofershtein and S. Cohen, "Predicting Email Recipients." ACM Press, 2015, pp. 761–764. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2808797.2808805

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

[11] A. Nordb, "Data Visualization for Discovery of Digital Evidence in Email," 2014. [Online]. Available: http://brage.bibsys.no/xmlui/handle/11250/198551

[12] K. K. Waterman and P. J. Bruening, "Big Data analytics: risks and responsibilities," *International Data Privacy Law*, vol. 4, no. 2, pp. 89–95, May 2014. [Online]. Available: http://idpl.oxfordjournals.org/content/4/2/89

[13] J. Shetty and J. Adibi, "The Enron email dataset database schema and brief statistical report," *Information sciences institute technical report, University of Southern California*, vol. 4, 2004. [Online]. Available: http://foreverdata.com/1009/Enron_Dataset_Report.pdf

[14] G. Tang, J. Pei, and W.-S. Luk, "Email mining: tasks, common techniques, and tools," *Knowledge and Information Systems*, vol. 41, no. 1, pp. 1–31, Jun. 2013. [Online]. Available: http://link.springer.com/10.1007/s10115-013-0658-2

[15] K. Yelupula and S. Ramaswamy, "Social network analysis for email classification," in *Proceedings of the 46th Annual Southeast Regional Conference on XX*. ACM, 2008, pp. 469–474. [Online]. Available: http://dl.acm.org/citation.cfm?id=1593229

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

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

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

[19] G. Wilson and W. 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*. IEEE, 2009, pp. 3256–3263. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4983357

[20] 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*. ACM, 2007, pp. 109–117. [Online]. Available: http://dl.acm.org/citation.cfm?id=1348562