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

Kayla M. Straub, Joseph M. Ernst, William C. Headley, Robert W. McGwier
Hume Center for National Security and Technology,
Department of ECE, Virginia Tech, Blacksburg, VA 24060, USA
Email: kstraub@vt.edu, jmernst@vt.edu, cheadley@vt.edu, rwmcgwi@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 released, researchers were able to provide 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 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 modeled behavior patterns and classified employees by job title with 96% accuracy. The algorithm performed very well not only on participants in the study but also on other members of the center who were connected to participants through email. Furthermore, the data revealed relationships not present in the center's formal operating structure. The result is an organic organizaiton chart which contains much fuller understanding of the center's internal structure than can be found in the official organization chart.**

I. Introduction

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

The study presented in this paper collected an appropriately anonymized email metadata dataset to demonstrate to what extent this metadata could be used to determine this organic organizational chart. The study utilized email metadata from 36 voluntary participants. This metadata is used not only to analyze the voluntary participants, but also to what extent the non-participant members of the organization can be characterized.

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 predict each employee's job title. Finally, a comparison is drawn between relationships displayed in the data and those depicted in the formal organizational chart.

This paper continues by discussing the related works in Section II. Section III presents the process of data collection and some statistics of the dataset. The features extracted from the data are described in Section IV, and the methods investigated using these features are covered in Section V. The results of the analysis are presented in Section VI. Section VII concludes the paper and presents opportunities for future work.

II. Related Works

Email is a pervasive medium for communication in modern society—particularly in the workplace. In 2015, there were over 2.6 billion email users [1]. Corporate email alone accounts for 54.7% of worldwide email traffic. In fact, the average business email user sends and receives a total of 112 emails per day. Retention of large email archives has become common practice as computer memory technology imporves and becomes increasingly affordable. This is a considerable amount of untapped information that could be leveraged to characterize employees within an organization.

Since the Enron email dataset was released in 2004 [2], it has been extensively researched on topics including spam classification [3]; email categorization [4]; and recipient prediction [5]. However, there are known

flaws and discrepancies with even the most recent versions of this dataset—ranging from misspelled email addresses [6] to duplicate emails [7].

The existing literature on analyzing social email behavior is mainly divided into two categories: traffic-based and social-based [8]. 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. For example, the work by Gilbert [9] showed that the wording used in an email could be used to make inferences about the corporate hierarchy. However, email text contains very sensitive infromation and is often unavailable for a general email study. Therefore, features based on email text are not considered in this work.

Using features extracted from email metadata alone, [10] 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 has been shown to successfully categorize email behavior [3].

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 a graphical layout [11]. A common feature used in social network analysis is betweenness centrality, which comes in several different flavors first developed by Freeman [12]. 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. In [13], a betweenness centrality algorithm was used to determine community structures within an organization. Other successful metrics were used in [14], which detected the most important email users within a corporate network without using betweenness as a feature. The features used instead were: degree, the number of edges connected to a node; density, the ratio of actual edges to the number of possible edges; and proximity prestige, the ratio of nodes that can reach a node i to the average distance from those nodes to i.

Instead of considering exclusively traffic-based or social-based analytics, these features can 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", which was used to rank Enron employees [15]. However, this work did not report any quantifiable results. The aim of this paper is to utilize both types of features in order to produce a measureable comparison between institution's organic chart and its official organizational chart.

III. 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. Due to difficulties with the Enron dataset, as described in Section II, this study uses a new dataset generated from volunteers in one of the university's centers.

In consideration of the inherent privacy concerns, researchers worked with the Internal Review Board (IRB) to approve a dataset which maintains participants' privacy. This dataset is meant to be representative of metadata which any company could use without violating the privacy of their employees. Special care was taken to protect the privacy of those involved in the study. During the collection process, all subject and body text was hashed using MD5, and all email metadata was stored in a secure database using scripts without any researchers observing any email text. Furthermore, any identifying information has been omitted from this publication. The following is the metadata provided from each email:

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

Table I compares statistics between this internal dataset and the Enron corpus. This internal database is more modern, contains more emails, and covers a longer time period. However, this study involved fewer people than were used to construct the Enron dataset. While the study includes only 36 volunteers, the email metadata from these volunteers identified 38 additional employees of the center. These additional employees were included in the study when ground truth for their job was available and when they were identified in at least 100 email metadata records. This provides 74 total employees in the study.

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 challenge

TABLE I
A COMPARISON BETWEEN THE INTERNAL DATASET AND THE ENRON EMAIL CORPUS.

	Center	Enron
Time	11/2012-11/2015	1/2000-9/2002
Distinct Addresses	32,118	75,406
Participants	36	158
Distinct Emails	585,096	252,759

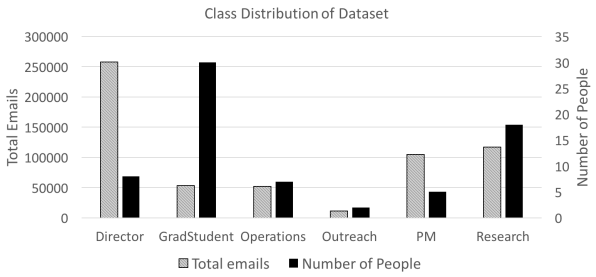

Fig. 1. Representation of each class in the database with respect to total number of emails and number of people. The number of people with a given job title is often not a reliable indicator of how many emails belonged to that class.

of working with this dataset is that the distribution of these job titles is far from uniform. For example, approximately half of the participants are graduate students, and there are only two outreach personnel included in the study. Figure 1 compares the distribution of job titles to the distribution of total emails in the dataset. Even though graduate students are by far the largest class, their emails only comprise about 10% of the dataset. The directors exhibit opposite behavior: there are only eight directors in the center, but collectively their emails make up almost 50% of the database.

IV. FEATURES

The study includes 114 features that were extracted from the email data: 84 traffic-based and 30 social-based. The traffic-based features focus on the amount and types of emails each employee sends and receives. Whereas thehe social features aim to quantify relationships within the community structure of the organization. In the following sections, all features from each of the two categories are described.

A. Traffic-Based Features

Traffic-based features include total number of emails, total sent, and total received. Other features are based on whether recipients were the primary recipient of the email or were copied on the email. Similarly, the number of emails sent and received as replies or forwards were used. Features also included the number of unique email address connections and the number of unique email subjects, sent and received. The average number of recipients on emails sent and received for each participant were calculated. 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.

Other useful information collected from emails included the timestamp, character counts of the subject and body, and the presence of any attachments. Utilizing the timestamp informaiton, the total number of emails sent and received after hours and the number of emails sent and received after hours within the center were also used as metrics. For this purpose, after hours was defined as between 6pm and 7am EST on weekdays or anytime on weekends. 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 number of attachments sent and received were computed as well as the average number of attachments sent and received per email. Digitally signing an email or encrypting an email are recognized as a type of attachment. Additional features were calculated from the total number of signed or encrypted emails sent and received as well as unique email addresses and subjects with signed or enctypted emails.

The features above in general involved raw email counts. To normalize some of the features, metrics were additionally calculated as a percentage. Examples include the percentage of emails that were sent after hours and the percentage of all received emails with unique subjects out of all received emails.

B. Social Network Features

In addition to tracking metadata statistics, features are also derived from modeling the emails as a social network. A social network is composed of nodes, which represent people, and edges, which represent the emails between people. For this analysis, two different graphs were generated. In the full graph, an edge exists between any two individuals that exchanged at least one email. A second graph only produces an edge between two nodes if at least 10 emails were exchanged, in order to focus on ongoing email relationships. A representation of the full graph is shown as an adjacency matrix in Figure 2. Each of the two axes represent the employees of the center. The color at each coordinate indicates the volume of emails sent between the two employees. Some employees never exchanged any emails, while others exchanged several thousand.

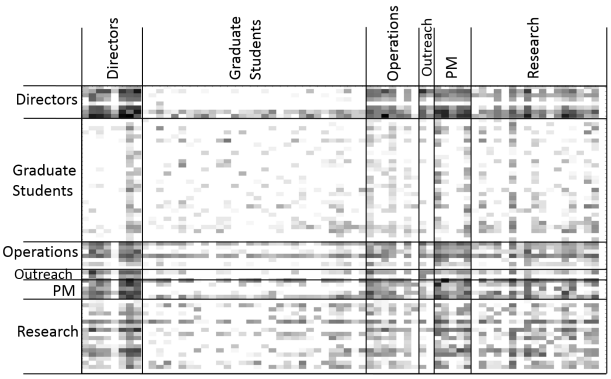

Fig. 2. The adjacency matrix representing the social connections of the center. This graph is very well connected with just one component. Nonetheless, there are groups who never exchanged a single email.

With this representation, statistics can be calculated about the people in the graph. The degree of each node was calculated for both the partial and full graph. The degree of a node i is the number of other nodes connected to node i, and these connected nodes are called neighbors of i. Additionally, the average neighbor degree of each node was computed as a feature. For a node i, this metric averages the degree of each node in the neighborhood of i, that is all nodes connected to i. The distances between nodes were also used to generate some features. The average shortest path metric calculates the length of the shortest paths between node i and all other nodes in the graph G, and it returns the average of these path lengths. Similarly, the maximum shortest path length, or eccentricity, was used as a feature in the learning algorithm.

Some of the social features were based on existing graph theory concepts and algorithms. If a subgraph of a graph G is maximally connected, that is all nodes are connected directly to each other, then this is called a clique. The number of cliques to which a node belongs was used as a feature. The hubs and authorities of each node in both graphs were calculated. The terms hubs and authorities come from the Hyperlink-Induced Topic Search (HITS algorithm) developed by [16]. This algorithm was originally designed to rate web pages, but has since been applied to social networks. A node's authority is just that—a measure of its importance over other nodes. A node's hub score is a measure of how well-connected it is to other nodes.

Another algorithm used to generate features was the pagerank algorithm, developed by Google [17] also to rank webpages for search results. The assumption is that the most important webpages will be linked to frequently by other pages. Therefore, the ranking is determined by estimating the quality and quantity of links to a node. The square clustering coefficient for each node was used as a feature. Say there exists a node i with neighbors j and k. This metric, developed by [18], measures the probability that j and k are also neighbors to a fourth node, l. The higher the clustering coefficient, the more connected the node is within its neighborhood. The triangle clustering coefficient was also used as a metric. This value, developed by [19], is the same as the square clustering coefficient but instead determines the probability of connected triangles involving each node.

The majority of the social-based features were centrality measures. This includes closeness centrality, betweenness centrality, degree centrality [20], current flow closeness centrality, current flow betweenness centrality [21], communicability centrality, communicability betweenness centrality [22], and load centrality [**?**].

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

V. Analysis

Due to the large number of features and relatively low number of participants, a classification method was carefully chosen to avoid overfitting the data. While tree based classifiers can be susceptible to overfitting, the random forest is robust to overfitting issues and was therefore chosen for this study. The java-based software package Weka was used to generate the random forest based on the algorithm described in [23].

Random forest training is an ensemble method of machine learning, comprised of many random trees. A random tree is an algorithm that uses training data to learn a series of rules to divide the data into the various classes. 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.

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

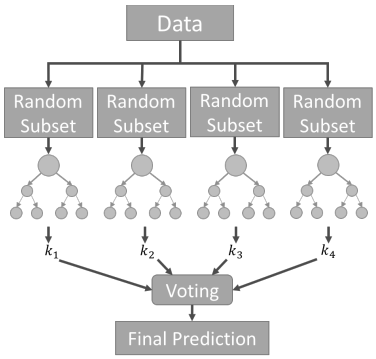

Fig. 3. The Random forest method aggregates many random trees via bagging to build a robust classifier.

for that tree. A validation set was used to determine the optimal values for the hyperparameters of the algorithm. This analysis determined that each forest should contain 750 trees and each tree would use a subsample of 7 random features.

After all the trees are built from the training data, the test data is run through all the random trees in the forest. Each tree outputs a prediction label for each data point, and the majority vote on each sample is the final predicted label. This random forest model reduces the variance and increases the accuracy of the model compared to a single random tree. A visualization of the random forest training process is shown in Figure 3.

Random forests can be difficult to interpret because the ensemble method obscures which features are most meaningful. A feature analysis helps to better understand which features are better label predictors. Since random trees use mutual information to optimize branch splits, mutual information was used as the evaluation criteria for the features. Specifically, each attribute was evaluated by measuring the mutual information with respect to the class. In this model, both the feature value and the class are treated as random variables. Mutual information represents how well knowledge of the attribute informs prediction of the class. Mutual information is calculated as follows:

$$I(\text{Class};\text{Feature}) = H(\text{Class}) - H(\text{Class}|\text{Feature}) \quad (1)$$

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

After this mutual information value was calculated for each feature, they were ranked in order of most important to least. Table II shows the top twenty features from this analysis and the features' corresponding mutual information. Highlighted row represent the features either unique to this new dataset or not used previously in email analysis.

TABLE II
TOP 20 FEATURES RANKED BY THE MUTUAL INFORMATION.

Feature	Type	Mutual Information
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 carbon 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 (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

VI. RESULTS

This section first shows the algorithm's ability to correctly classify both the study's volunteers and the additional employees identified from the volunteers' emails. The second part of the results section assumes perfect labeling of the employees and compares email interactions to the official organizational chart. The ultimate goal of this research was to determine the organic organizational chart, generated entirely from the email data.

A. Classification Results

The data was split by randomly assigning each email to either the training or testing set with equal probability. Then, all of the metrics described in Section IV were calculated for both groups separately. The training data was used as input to the random forest algorithm as described in Section V, and predictions were generated for the test data. The number of correct and incorrect classifications for each class are shown below in Figure 4. Five of the 37 participants in the study were very new employees to the center, and were not considered in this analysis due to lack of sufficient email data (defined as less than 100 total emails). Note that only three predictions were wrong, and all were a result of confusing research staff with graduate students. It is important to note that two of these misclassifications are for employees who did not provide their emails for the study. Therefore, the classification accuracy for the study participants is 96.8%, correctly classified inferred employees is 94.7%, and the overall accuracy of this method using all features is 95.7%.

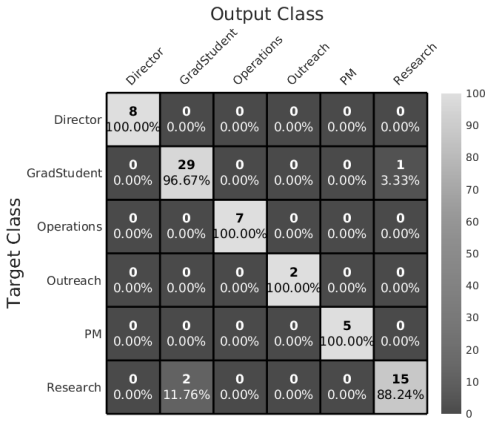

Fig. 4. The Random Forest algorithm was extremely accurate even for very uneven class sizes. Note that all members of the 4 smallest classes were labeled perfectly. There were only 3 errors out of 69 employees, two of which for employees who did not provide emails for the study.

B. Hierarchy Analysis

Most of the employees at the center are organized under a director and work with a program manager (unless for example they are a director or program manager). To generate a metric of how well emails can be used to predict the center's organizational chart, the director and project manager for each applicable employee is predicted from the email metadata.

The director of each employee is predicted by the algorithm to be the director that the employee communicated with most by email. Only 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 most frequently with their primary program manager. The relation between employees to project managers appears to be stronger than that with directors. Many of the errors in this classification are due to employees who work with multiple project managers.

These statistics are depicted for all classes in Figure 5. Note that employees with job titles of project manager, outreach, and operations have directors, but no equivalent to project manager.

Finally, the predicted labels from Section VI-A were combined email patterns to generate the organic hierarchy in Figure 6. This chart represents the flow of project information for the center: from directors to project managers to research staff down to graduate students. Outreach and operations personnel are not

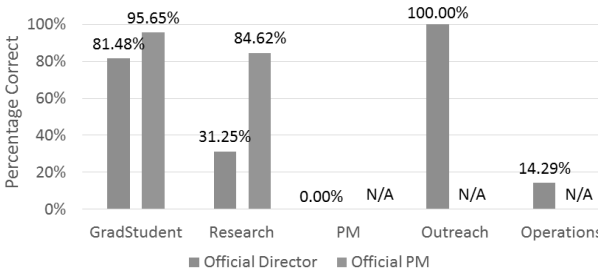

Fig. 5. The official director an employee, from the organization chart, is often not the director with whom they exchange the most emails. However, graduate students and researchers often communicate most

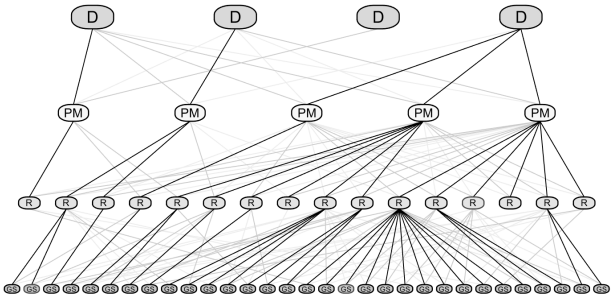

Fig. 6. Project organization chart generated from the email data. This analysis could be used in corporate merges or restructurings in order to better understand the communication network of an organization.

actively involved in project work, and are therefore omitted from this chart. Edges were draw from one layer to the layer above based on how many emails were sent. For example, each graduate student has three edges corresponding to each of the three researchers they emailed most frequently. The most common researcher is the darkest line, the second-most common edge is more transparent, and the third-most common link is the most transparent.

VII. Conclusions and Future Work

This work presents a new dataset, larger than Enron, that was collected from volunteers' emails with particular attention to protect participant privacy. The new dataset includes accurate labels executed by researchers with knowledge of the center and its employees. Statistics were calculated from this dataset, and were used with a random forest algorithm to automatically classify the center's employees. Random Forests are shown to be powerful classifiers by predicting employee job titles with very high accuracy, even for employees for whom only secondhand data is available in the dataset. The email data was also used to show that emails could be used to predict an employee's primary program manager, but had a worse chance of being able to identify the director associated with the employee on the official organizational chart. This work has shown that it is pos-

sible to generate important organizational information from using carefully processed email metadata without compromising the privacy of employees. These methods could be applied to organizational analysis, anomalous behvaior detection, leadership identification of any communications system. Future work will attempt to further explore the center's organic hierarchy through this data as well as to apply these algorithms to other datasets to determine the general applicability of the results.

REFERENCES

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

[2] B. Klimt and Y. Yang, "Introducing the enron corpus." in *CEAS*, 2004.

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

[4] 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*. IEEE, 2014, pp. 269–272.

[5] 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*. ACM, 2015, pp. 761–764.

[6] A. Nordbø, "Data visualization for discovery of digital evidence in email," 2014.

[7] K. K. Waterman and P. J. Bruening, "Big data analytics: risks and responsibilities," *International Data Privacy Law*, vol. 4, no. 2, pp. 89–95, 2014.

[8] 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, 2014.

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

[10] 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.

[11] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.

[12] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, pp. 35–41, 1977.

[13] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, "E-mail as spectroscopy: Automated discovery of community structure within organizations," *The Information Society*, vol. 21, no. 2, pp. 143–153, 2005.

[14] 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.

[15] 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.

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

[17] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web." 1999.

[18] P. G. Lind, M. C. González, and H. J. Herrmann, "Cycles and clustering in bipartite networks," *Physical review E*, vol. 72, no. 5, p. 056127, 2005.

[19] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, and J. Kertesz, "Generalizations of the clustering coefficient to weighted complex networks," *Physical Review E*, vol. 75, no. 2, p. 027105, 2007.

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

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

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

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