Determining Formal and Informal Organizational Hierarchy

J.M. Brown, E.A. Benagh, C.G. Fournelle
Boston Fusion, Corp., 1 Van de Graaff Drive, Burlington, MA

***Abstract*—*Social hierarchy influences how information flows within groups of people, but it is not obvious how interactions within social networks reflect or differ from formal hierarchies and how well one can be used to predict the other. Using records from a business unit of ~6,000 employees, we determined formal organizational structure from the LDAP (Lightweight Directory Access Protocol) and informal structure from email communications over the course of one year. We compared features within the email network with relationships in the formal hierarchy. We then trained a SVM classifier to predict both supervisor-supervisee pairs and pairs within the same LDAP group based on email network features. The relationship from supervisee to supervisor was reflected most strongly in the email features. In all cases, the classifier found strong false positives which may reflect the cross-matrix structure of this organization and indicate mentoring or team relationships not included in the formal structure.*

Keywords*—*graph analysis, social network, machine learning, hierarchy determination

1. Introduction

Hierarchy is an integral component of human social organization, affecting how people relate to each other, how information travels within groups, and how organizations evolve over time. In organizations, hierarchies are often formally defined with supervisors, groups, and clear chains of command. However, this formal hierarchy may not reflect the actual working hierarchy of people's day-to-day interactions. Social communication networks, such as emails, reveal these day-to-day interactions but carry no explicit record of hierarchy. If we can determine the underlying connection between social networks and formal hierarchies, we might be able to infer one from the other. Further, differences between these two networks reveal discrepancies between the true organizational structure and the "effective" organizational structure.

Prior research examined social networks with respect to organizational hierarchy. In [8], the authors propose an automated detection algorithm for social hierarchy by developing a 'social score' for each user and comparing user scores to determine their rank within the levels of the organization. The algorithm demonstrated success for recognizing high levels of the organization but struggled with the lower echelons. Likewise, [9] generated a ranked list of node importance based on an entropy model. An algorithm for determining the type of organizational structure is presented in [6]. Many studies have been limited by difficulty in finding real-world email data. Several papers (e.g., [4, 8]) focus on the Enron email corpus containing ~600,000 messages between 158 senior employees (e.g., [10]).

In this paper, we focus on characterizing the links between individuals to identify the nature of the relationship. The premise is that interactions with supervisors and group members have distinct characteristics regardless of organizational level. If we can determine these small-scale connections, then we can infer the large-scale organizational structure by aggregating the connections. We use records from a business unit of ~6,000 people to provide the large volume of data suited to a machine learning approach. In the sections that follow, we discuss the data collection, individual features in the email network with respect to the organizational hierarchy, and a support vector machine (SVM) classifier for predicting formal relationships from the email communications. We then present and evaluate our results, summarize our findings, and discuss ideas for future improvements.

2. Data

In order to test the effects of hierarchy, we used real-world information for employees of a large company. All computer activity for the ~6,000 employees was tracked over the course of a year from January 1 to December 31, 2014. Table 1 contains summary statistics of the data properties. Due to the sensitive nature of the data collected, limited information was available for export and exported information was carefully anonymized.

The company maintained an LDAP (Lightweight Directory Access Protocol) for all employees to manage information access, providing a definitive guide to the formal organizational hierarchy. We recorded daily changes in the LDAP files to identify events such as an employee changing supervisor, moving departments, and joining or leaving the organization. The company hand curated the LDAP; hence, events may be offset from the record by up to a couple of days.

The email communications provide an extensive view of the social network within the organization and form the basis for building the informal organization structure. We have no record of email content due to the private nature of the information. Email features include basic metadata of the email address that was used to send the message, all recipients (as email addresses), when the message was sent, the length of the subject line, and the size of the contents

and attachments. All email addresses were consistently anonymized to enable researchers to reconstruct the structure of the communication network without revealing the real identities of individuals that were using the email accounts. Some users in the LDAP do not appear in the email network, either due to a lack of records or difficulty reconstructing the email address to user connections stripped in the anonymization process.

Table 1. Properties of the LDAP and email data between Jan. 1, 2014 and Dec. 31, 2014

LDAP properties	
#LDAP Employees	6,562
#Departments	954
#Supervisors	827
#Groups	1,075

Email properties	
#Employees	4,794
#Connections	260,318
#Total emails	3.8M
#Email addresses	36,811
#Supervisor links	4,030
#Group links	36,861

3. Identifying Formal Hierarchy

To determine the formal hierarchy, we identified supervisor-supervisee pairs within the LDAP. To protect company proprietary information, some higher level employees were excluded from the sample for which email data was collected. We identified all pairs of employees that belong to a common group, where we define a group as employees that share a supervisor and the supervisor themselves. Group sizes ranged from 2 to 698 with a median of 5. The structure changed daily (although not significantly) as people were hired, quit, or changed supervisors and departments.

Figure 1 shows the organization hierarchy on a supervisor group basis (i.e., with each node representing a group, and a link indicating communications between the associated groups). We canonically identify each group node with the supervisor, and we display node size in proportion to group size. When more than one supervisor relationship existed over the time period, the figure shows the longest lasting relationship to provide an overview of the structure. The highest level supervisor is generally not tracked in the email sample (shown in the figure in blue), due to the corporate sensitivity of high level executive work. Within the data set, there are two large hierarchy structures and one group of several hundred people who have no additional subordinates within the group. Overall, there are 28 structured multi-group components plus an additional ~50 supervisor groups led by supervisors who are excluded from the email sample.

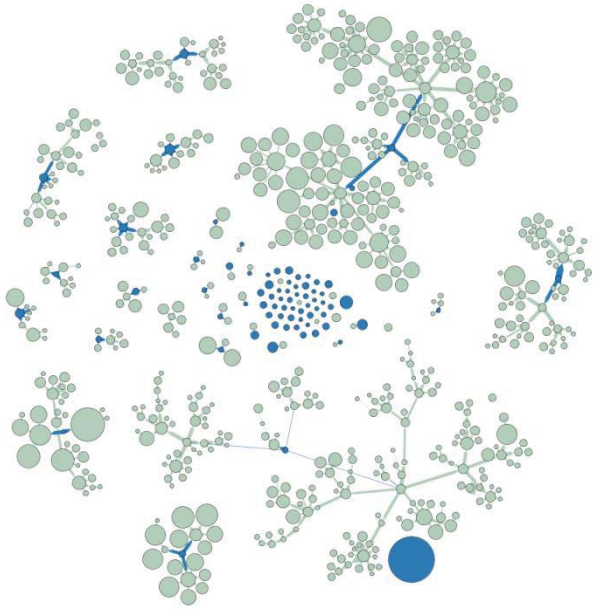

Figure 1. The organizational structure of supervisor groups. Every node represents a supervisor group with the size proportional to the size of the group. Light green nodes are groups that are led by individuals within the experimental sample, while blue nodes are led by supervisors outside of the email sample. Edge weights show the fraction of the year the connection existed.

4. Identifying Informal Network

To construct the communication network among employees, we first connected all the anonymized email addresses to individual employees. Employees generally had several addresses and more than one mail application, with each application recording email addresses in slightly different formats. The anonymization process then stripped the original information, making it more challenging to associate differently formatted addresses. The dataset also includes activity for group email addresses with large distribution lists that were used by multiple people.

The email tracking software recorded two types of email events: *send* events, when the user sent an email, and *view* events, when the user viewed an email, which included initial viewing of email messages as well as rereading events. We created a dictionary linking all user IDs to email address variants in two steps. First, we linked all the sender addresses in the send events with the user ID associated with the event. We then examined all the view events to identify other email variations. We dismissed view events that had more than one recipient, as it was too difficult to determine which of the addresses belonged to the individual associated with the event. We also required that either the

sender or recipient address was already in the dictionary to determine whether the user ID was associated with the sender or recipient address. If both addresses were already in the dictionary, we added the recipient address to a list of group emails sent to multiple recipients and removed the address from the dictionary. The dictionary clearly had imperfect performance due to limitations in the data, finding that only ~50% of supervisor-supervisee pairs exhibited email links. Some relationships were short term, on the order of days and weeks, and it is possible no emails were exchanged during those transitions.

Once the dictionary was as complete as possible, we captured graphs on a monthly basis. Each node in the graph represents a user, and the directed connections indicate that an email was sent. Each connection has a weight equal to the number of emails sent. Since single emails could be counted multiple times for each read-through, it was infeasible to accurately interpret the 'viewed event' records as weighted communication links. As a result, we chose to ignore the 'view events', and focused on the cleaner 'sent events'. Because supervisor and group relationships change with time, we excluded months where the relationship was not in place when generating email features.

5. Comparison of individual features

We examined five features to identify traits that might support inference about supervisor-supervisee or group relationships. First, we examined whether email volume was higher between supervisor-supervisee and other intra-group pairs. Figure 2 shows the results for the intra-group pairs. When low total numbers of emails were exchanged, the pairs overwhelmingly crossed group boundaries. If a pair had sent hundreds of emails, over half of those pairs are within the group.

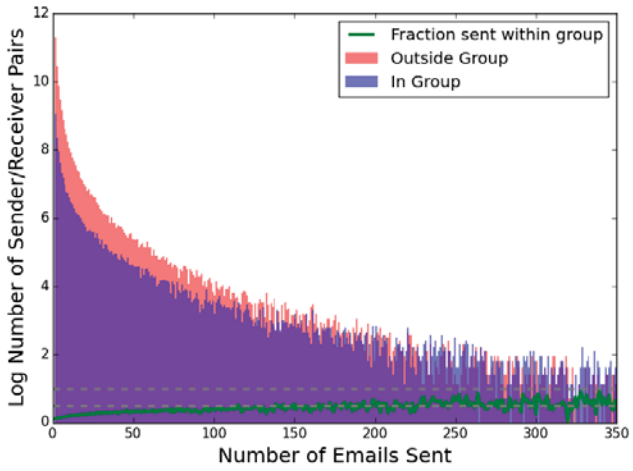

Figure 2. Comparison of emails sent within groups versus across group boundaries. The colored bars represent how many pairs received that number of emails. Note that the 'in group' and 'outside group' bars are overlaid rather than stacked. The green line shows the fraction of pairs which were within group for each number of emails sent. The two horizontal dashed grey lines mark 0.5 and 1.

The second feature we examined was the rank within the contact list. For each pair, we generated two rank scores: the first score indicates the position in the second's contact list, ordered by volume of emails, and the second score indicates the position of the second individual in the first's contact list.

As shown in Figure 3, employees generally have their supervisor as the first or second person in their contact list. Interestingly, this observation held at all levels of the organizational hierarchy. We divided the sample into those with direct reports and those without and saw little difference in the resulting distribution of rank scores. Supervisors generally contact their supervisees more frequently than others, but the effect is more diluted due to supervisors having multiple supervisees.

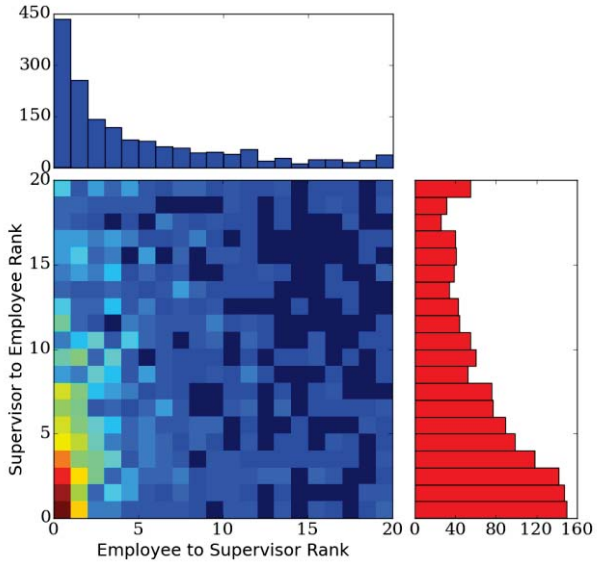

Figure 3. The supervisor to employee rank is the position of the employee in the supervisor's list of contacts ordered by number of emails sent, while the employee to supervisor rank in the supervisors position in the employee's list. The heatmap shows the 2-dimensional histogram of the two ranks, while the 1-dimensional histograms on the edges show the collapsed histogram in each direction.

Finally, we examined the similarity of the communication patterns of the pair. We found all the maximal cliques within the email graph using the algorithm first proposed by [2] and implemented by [1]. We then counted the number of cliques within which a pair co-occurs as a measure of overlapping circles of contacts. We observed that most pairs are not in group, but the ratio of in-group to out-of-group

increases with number of clique memberships, as shown in Figure 4.

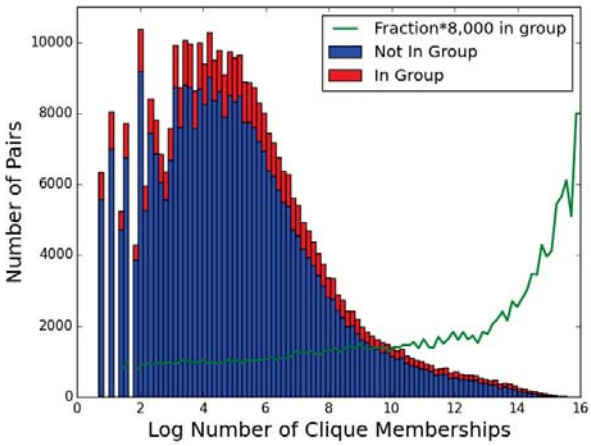

Figure 4. Histogram of the number of maximal cliques in which both individuals in a pair has membership in the same group. The histogram is stacked with the 'in group' population appears on top of the 'outside group'. The green line shows the fraction in group, multiplied by 8000 to align the scales.

For each individual in every pair of users, we computed the fraction of the user's total email volume that was sent to the other individual in that pair. The final feature was the mean size of the emails sent based on the number of characters in the email body to further characterize the significance of their interaction.

6. SVM Classifier

We used the support vector classifier (SVC) in the Python library 'scikit-learn' to combine all the individual features to produce better classification. We generated two classifiers, one to identify supervisor-supervisee from other relationships, and another to classify relationships as internal to and across group boundaries. We normalized all features to have a mean of 0 and a standard deviation of ±1 so that the different scales for the features did not influence the classification. The classes are highly unbalanced, with most relationships being neither supervisor-supervisee nor within group. In fact, supervisor links comprise only 1.5% of the connections, whereas group connections account for 14% of the total. To compensate, we weighted the classes by the inverse of frequency. The dataset was large enough that the classifiers still included instances of the underrepresented classes within the training sample. We explored using both linear and radial basis function (RBF) kernels, but achieved slightly better performance with the linear kernel. The training sample consisted of 50,000 instances with ~210,000 instances in the target population. Increasing the training size to 100,000 did not improve performance while greatly increasing runtime.

There are some limitations for classification over this set of data. First, the email structure (unsurprisingly) does not perfectly reflect the LDAP hierarchy. Second, the classifier can only evaluate relationships present in email, but some supervisors never email their supervisee, and vice versa. Finally, we are aware that the management for this organization is "matrixed", and individuals receive work assignments under managers outside of their formal LDAP structure. We therefore do not expect the classifiers to be able to achieve perfect performance.

7. Results

The classifier achieved meaningful classification of both supervisor and group relationships. Figure 5 shows the receiver operating characteristic (ROC) curve for both the supervisor-supervisee classifier and the group classifier. The area under the curve (AUC) for the supervisor classifier was 0.79 while the AUC for the group classifier was 0.72. The supervisor classifier had better performance than the group classifier, indicating more identifiable relationships between supervisor and supervisee than within groups. The group classifier is likely weakened by group members who have some communication but are not strongly tied. There is no indication that all members of a supervisor group actively worked together, and in the case of the largest 698 member group, it seems highly unlikely. That said, enough group members are connected strongly, and so the classifier can achieve meaningful performance.

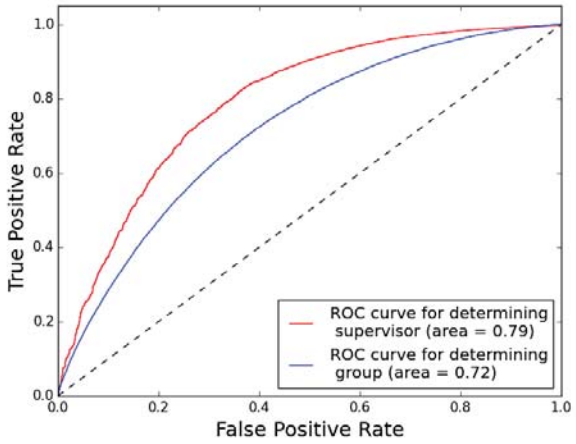

Figure 5. ROC curves for determining supervisor (red) and group membership (blue).

8. Evaluation

We present the optimal prediction results for both classifiers in Table 2. We determined the optimal threshold by maximizing the distance from the random chance line.

Of particular interest are the false positives identified by the classifiers. These may represent relationships that mimic

those of supervisor-supervisee and group (such as team leaders and teams working on the same project), but are not connected in the formal hierarchy, partly due to the matrixed organization style. We have no method to check these work arrangements in this dataset, but we examined these pairs and predicted supervisors in more detail. Of the false positives from the supervisor classifier, ~12,000 are group relationships but not supervisor relationships. Of all the false positives, 90% of the pairs represent 1,969 individuals who are incorrectly predicted as supervisors in five or more pairs. As the number of instances in which a person is predicted as a supervisor increases, it becomes more and more likely that the person plays a key role in the informal hierarchy. The average number of predictions identifying an individual as a supervisor was 9 and the largest number of predictions was 144 false positive pairs.

Table 2. Classifier performance results for the test portion of the data set

Supervisor	Predicted Supervisor	Predicted Not
Actual Supervisor	*1,444*	*832*
Actual Not Supervisor	*36,367*	*171,735*

Group	Predicted Group	Predicted Not
Actual Group	*20,242*	*11,930*
Actual Not Group	*54,227*	*123,979*

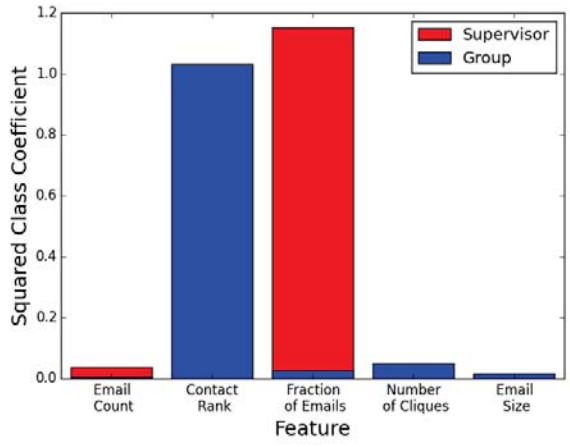

Figure 6. The feature importance for the two classifiers. Both classifiers have strong dependencies on single features.

To understand the importance of different features, we used the squared weight coefficients for each class, as determined by the linear classifier (as in [6]). Interestingly, both classifiers seem to strongly favor a single feature, but they favor different features. The supervisor classifier depends heavily on the fraction of an individual's messages that are sent to the supervisor, while the group classifier mainly uses contact rank.

9. Next Steps

In the future, we plan to utilize the insights into organizational hierarchy we gained from the email network to identify information flows through the network and to search for anomalies in communication. We will investigate the strong false positive to look for connections indicating work team relationships. We will also expand the feature sets for the classifiers to include features such as frequency of contact, response time, and number of attachments sent.

The technology we are developing will benefit operational customers who are trying to build an understanding of information and influence networks. Directed social networks are present throughout modern life, including social media such as Twitter, Facebook, and LinkedIn, as well as email, SMS, and telecommunication networks. Unlike companies, these social networks rarely come with a blueprint to the underlying hierarchical structure. It is only by inferring the hierarchy that we can understand the interpersonal connections. The relationship classifiers we developed here provide a method to locally identify rank within the large graph rather than globally trying to rank individuals.

10. Summary

We present a study of the organizational hierarchy of a corporate business unit of ~6000 people over the course of a year. We examined both the formal and informal hierarchy to determine how closely connected the two were and developed a classifier to predict formal relationships from the informal email network.

We found significant differences between the formal and informal organizational structure. Formal relationships may have no equivalent in the informal structure or may be less intense than expected. The informal structure includes many connections with no formal counterparts, as might be expected. Formal supervisor and group relationships, however, were noticeably different from other connections within the email network.

We examined five features in the email network for their relationship to the formal hierarchy: email volume, contact rank, fraction of user emails sent to contact, number of joint maximal cliques and email size. We then used a support vector machine classifier with a linear kernel to simultaneously use all the features for classification. Classification of supervisor relationships had an area under

the ROC curve of 0.79 while group classification had slightly poorer performance at 0.72. The email network thus reflects supervisor-supervisee relationships more strongly than group structure, although supervisor relationships are highly underrepresented in the sample as a whole.

The results from the classifier hint at a parallel informal structure that mimics the formal hierarchy. Strong false positives from the classifier may identify relationships that are functionally similar to supervisor and group, such as team leader and team. Identifying the actual organizational structure from communications provides a more complete organizational picture than depending on the incomplete formal hierarchy.

11.Acknowledgements

The authors gratefully acknowledge support for this work from DARPA through the ADAMS (Anomaly Detection At Multiple Scales) program for funding project GLAD-PC (Graph Learning for Anomaly Detection using Psychological Context), under U.S. Army contract number W911NF-11-C-0216. We gratefully acknowledge the support of Xerox PARC and our teammates there including Dave Gunning, Gaurang Gavai, Sricharan Kumar, and John Hanley. Any opinions, findings, and conclusions or recommendations in this material are those of the authors and do not necessarily reflect the views of the government funding agencies.

12. References

1. Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, "Exploring Network Structure, Dynamics, and Function Using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008.
2. Bron, C. and J. Kerbosch, "Algorithm 457: Finding All Cliques of an Undirected Graph", Commun. ACM 16, 9, 575-577, 1973.
3. Clauset, A., M.E.J. Newman, C. Moore, "Finding Community Structure in Very Large Networks", Phys. Rev. E (70), 066111, 2004.
4. Diesner, J., T.L. Frantz, K.M. Carley, "Communication Networks from the Enron Email Corpus", Computational & Mathematical Organization Theory, 11, 3, 201, 2005.
5. Gupte, M., P. Shankar, et al. "Finding Hierarchy in Directed Online Social Networks", WWW 2011, Mar. 28 – Apr 1, 2001, Hyperabad, India, 557, 2001.
6. Guyon, I., J. Weston, S. Barnhill, V. Vapnik. "Gene Selection for Classification using Support Vector Machines", Machine Learning, 46, 389, 2002
7. Maiya, Arun S., Tanya Y. Berger-Wolf, "Inferring the Maximum Likelihood Hierarchy in Social Networks", Proceeding of CSE '09 (4), 245, 2009.
8. Rowe, R., G. Creamer, S.Hershkop, S. Stolfo, "Automated Social Hierarchy Detection through Email Analysis", Joint 9th WEBKDD & 1st SNA-KDD, San Jose, CA, Aug. 12, 2007
9. Shetty, J. and J. Adibi, "Discovering Important Nodes through Graph Entropy: The Case of Enron Email Database," LinkKDD '05, ACM, 74, 2005
10. http://www.cs.cmu.edu/~enron/