A Comprehensive Gold Standard for the Enron Organizational Hierarchy

Apoorv Agarwal[1]* **Adinoyi Omuya[1]**** **Aaron Harnly[2]†** **Owen Rambow[3]‡**

[1] Department of Computer Science, Columbia University, New York, NY, USA
[2] Wireless Generation Inc., Brooklyn, NY, USA
[3] Center for Computational Learning Systems, Columbia University, New York, NY, USA
* apoorv@cs.columbia.edu ** awo2108@columbia.edu
†aaron@cs.columbia.edu ‡rambow@ccls.columbia.edu

Abstract

Many researchers have attempted to predict the Enron corporate hierarchy from the data. This work, however, has been hampered by a lack of data. We present a new, large, and freely available gold-standard hierarchy. Using our new gold standard, we show that a simple lower bound for social network-based systems outperforms an upper bound on the approach taken by current NLP systems.

1 Introduction

Since the release of the Enron email corpus, many researchers have attempted to predict the Enron corporate hierarchy from the email data. This work, however, has been hampered by a lack of data about the organizational hierarchy. Most researchers have used the job titles assembled by (Shetty and Adibi, 2004), and then have attempted to predict the relative ranking of two people's job titles (Rowe et al., 2007; Palus et al., 2011). A major limitation of the list compiled by Shetty and Adibi (2004) is that it only covers those "core" employees for whom the complete email inboxes are available in the Enron dataset. However, it is also interesting to determine whether we can predict the hierarchy of other employees, for whom we only have an incomplete set of emails (those that they sent to or received from the core employees). This is difficult in particular because there are dominance relations between two employees such that no email between them is available in the Enron data set. The difficulties with the existing data have meant that researchers have either not performed quantitative analyses (Rowe et al., 2007), or have performed them on very small sets: for example, (Bramsen et al., 2011a) use 142 dominance pairs for training and testing.

We present a new resource (Section 3). It is a large gold-standard hierarchy, which we extracted manually from pdf files. Our gold standard contains 1,518 employees, and 13,724 dominance pairs (pairs of employees such that the first dominates the second in the hierarchy, not necessarily immediately). All of the employees in the hierarchy are email correspondents on the Enron email database, though obviously many are not from the core group of about 158 Enron employees for which we have the complete inbox. The hierarchy is linked to a threaded representation of the Enron corpus using shared IDs for the employees who are participants in the email conversation. The resource is available as a MongoDB database.

We show the usefulness of this resource by investigating a simple predictor for hierarchy based on social network analysis (SNA), namely degree centrality of the social network induced by the email correspondence (Section 4). We call this a lower bound for SNA-based systems because we are only using a single simple metric (degree centrality) to establish dominance. Degree centrality is one of the features used by Rowe et al. (2007), but they did not perform a quantitative evaluation, and to our knowledge there are no published experiments using only degree centrality. Current systems using natural language processing (NLP) are restricted to making informed predictions on dominance pairs for which email exchange is available. We show (Section 5) that the upper bound performance of such

NLP-based systems is much lower than our SNA-based system on the entire gold standard. We also contrast the simple SN-based system with a specific NLP system based on (Gilbert, 2012), and show that even if we restrict ourselves to pairs for which email exchange is available, our simple SNA-based systems outperforms the NLP-based system.

2 Work on Enron Hierarchy Prediction

The Enron email corpus was introduced by Klimt and Yang (2004). Since then numerous researchers have analyzed the network formed by connecting people with email exchange links (Diesner et al., 2005; Shetty and Adibi, 2004; Namata et al., 2007; Rowe et al., 2007; Diehl et al., 2007; Creamer et al., 2009). Rowe et al. (2007) use the email exchange network (and other features) to predict the dominance relations between people in the Enron email corpus. They however do not present a quantitative evaluation.

Bramsen et al. (2011b) and Gilbert (2012) present NLP based models to predict dominance relations between Enron employees. Neither the test-set nor the system of Bramsen et al. (2011b) is publicly available. Therefore, we compare our baseline SNA based system with that of Gilbert (2012). Gilbert (2012) produce training and test data as follows: an email message is labeled *upward* only when every recipient outranks the sender. An email message is labeled *not-upward* only when every recipient does not outrank the sender. They use an n-gram based model with Support Vector Machines (SVM) to predict if an email is of class *upward* or *not-upward*. They make the phrases (n-grams) used by their best performing system publicly available. We use their n-grams with SVM to predict dominance relations of employees in our gold standard and show that a simple SNA based approach outperforms this baseline. Moreover, Gilbert (2012) exploit dominance relations of only 132 people in the Enron corpus for creating their training and test data. Our gold standard has dominance relations for 1518 Enron employees.

3 The Enron Hierarchy Gold Standard

Klimt and Yang (2004) introduced the Enron email corpus. They reported a total of 619,446 emails taken from folders of 158 employees of the Enron corporation. We created a database of organizational hierarchy relations by studying the original Enron organizational charts. We discovered these charts by performing a manual, random survey of a few hundred emails, looking for explicit indications of hierarchy. We found a few documents with organizational charts, which were always either Excel or Visio files. We then searched all remaining emails for attachments of the same filetype, and exhaustively examined those with additional org charts. We then manually transcribed the information contained in all org charts we found.

Our resulting gold standard has a total of 1518 nodes (employees) which are described as being in immediate dominance relations (manager-subordinate). There are 2155 immediate dominance relations spread over 65 levels of dominance (CEO, manager, trader etc.) From these relations, we formed the transitive closure and obtained 13,724 hierarchal relations. For example, if A immediately dominates B and B immediately dominates C, then the set of valid organizational dominance relations are A dominates B, B dominates C and A dominates C. This data set is much larger than any other data set used in the literature for the sake of predicting organizational hierarchy.

We link this representation of the hierarchy to the threaded Enron corpus created by Yeh and Harnley (2006). They pre-processed the dataset by combining emails into threads and restoring some missing emails from their quoted form in other emails. They also co-referenced multiple email addresses belonging to one person, and assigned unique identifiers and names to persons. Therefore, each person is a-priori associated with a set of email addresses and names (or name variants), but has only one unique identifier. Our corpus contains 279,844 email messages. These messages belong to 93,421 unique persons. We use these unique identifiers to express our gold hierarchy. This means that we can easily retrieve all emails associated with people in our gold hierarchy, and we can easily determine the hierarchical relation between the sender and receivers of any email.

The whole set of person nodes is divided into two parts: **core** and **non-core**. The set of core people are those whose inboxes were taken to create the Enron

email network (a set of 158 people). The set of non-core people are the remaining people in the network who either send an email to and/or receive an email from a member of the core group. As expected, the email exchange network (the network induced from the emails) is densest among core people (density of 20.997% in the email exchange network), and much less dense among the non-core people (density of 0.008%).

Our data base is freely available as a MongoDB database, which can easily be interfaced with using APIs in various programming languages. For information about how to obtain the database, please contact the authors.

4 A Hierarchy Predictor Based on the Social Network

We construct the **email exchange network** as follows. This network is represented as an undirected weighted graph. The nodes are all the unique employees. We add a link between two employees if one sends at least one email to the other (who can be a TO, CC, or BCC recipient). The weight is the number of emails exchanged between the two. Our email exchange network consists of 407,095 weighted links and 93,421 nodes.

Our algorithm for predicting the dominance relation using social network analysis metric is simple. We calculate the degree centrality of every node in the email exchange network, and then rank the nodes by their degree centrality. Recall that the degree centrality is the proportion of nodes in the network with which a node is connected. (We also tried eigenvalue centrality, but this performed worse. For a discussion of the use of degree centrality as a valid indication of importance of nodes in a network, see (Chuah and Coman, 2009).) Let $C_D(n)$ be the degree centrality of node n, and let DOM be the dominance relation (transitive, not symmetric) induced by the organizational hierarchy. We then simply assume that for two people p_1 and p_2, if $C_D(p_1) > C_D(p_2)$, then DOM(p_1,p_2). For every pair of people who are related with an organizational dominance relation in the gold standard, we then predict which person dominates the other. Note that we do not predict if two people are in a dominance relation to begin with. The task of predicting if two people are in a dominance relation is different and we do not address that task in this paper. Therefore, we restrict our evaluation to pairs of people (p_1, p_2) who are related hierarchically (i.e., either DOM(p_1,p_2) or DOM(p_2,p_1) in the gold standard). Since we only predict the directionality of the dominance relation of people given they are in a hierarchical relation,[1] the random baseline for our task performs at 50%. We have 13,724 such pairs of people in the gold standard. When we use the network induced simply by the email exchanges, we get a remarkably high accuracy of 83.88% (Table 1). We denote this system by SNA_G.

Type	# pairs	%Acc
All	13,724	83.88
Core	440	79.31
Inter	6436	93.75
Non-Core	6847	74.57

Table 1: Prediction accuracy by type of predicted organizational dominance pair; "Inter" means that one element of the pair is from the core and the other is not; a negative error reduction indicates an increase in error

In this paper, we also make an observation crucial for the task of hierarchy prediction, based on the distinction between the core and the non-core groups (see Section 3). This distinction is crucial for this task since by definition the degree centrality measure (which depends on how accurately the underlying network expresses the communication network) suffers from missing email messages (for the non-core group). Our results in table 1 confirm this intuition. Since we have a richer network for the core group, degree centrality is a better predictor for this group than for the non-core group.

We also note that the prediction accuracy is by far the highest for the **inter** hierarchal pairs. The inter hierarchal pairs are those in which one node is from the core group of people and the other node is from the non-core group of people. This is explained by the fact that the core group was chosen by law enforcement because they were most likely to contain information relevant to the legal proceedings against Enron; i.e., the owners of the mailboxes

[1] This style of evaluation is common (Diehl et al., 2007; Bramsen et al., 2011b).

were more likely more highly placed in the hierarchy. Furthermore, because of the network characteristics described above (a relatively dense network), the core people are also more likely to have a high centrality degree, as compared to the non-core people. Therefore, the correlation between centrality degree and hierarchal dominance will be high.

5 Using NLP and SNA

In this section we compare and contrast the performance of NLP-based systems with that of SNA-based systems on the Enron hierarchy gold standard we introduce in this paper. This gold standard allows us to notice an important limitation of the NLP-based systems (for this task) in comparison to SNA-based systems in that the NLP-based systems require communication links between people to make a prediction about their dominance relation, whereas an SNA-based system may predict dominance relations without this requirement.

Table 2 presents the results for four experiments. We first determine an upper bound for current NLP-based systems. Current NLP-based systems predict dominance relations between a pair of people by using the language used in email exchanges between these people; if there is no email exchange, such methods cannot make a prediction. Let G be the set of all dominance relations in the gold standard ($|G| = 13,723$). We define $T \subset G$ to be the set of pairs in the gold standard such that the people involved in the pair in T communicate with each other. These are precisely the dominance relations in the gold standard which can be established using a current NLP-based approach. The number of such pairs is $|T| = 2,640$. Therefore, if we consider a perfect NLP system that correctly predicts the dominance of $2,640$ tuples and randomly guesses the dominance relation of the remaining $11,084$ tuples, the system would achieve an accuracy of $(2640 + 11084/2)/13724 = 59.61\%$. We refer to this number as the upper bound on the best performing NLP system for the gold standard. This upper bound of 59.61% for an NLP-based system is lower (24.27% absolute) than a simple SNA-based system (SNA_G, explained in section 4) that predicts the dominance relation for all the tuples in the gold standard G.

As explained in section 2, we use the phrases provided by Gilbert (2012) to build an NLP-based model for predicting dominance relations of tuples in set $T \subset G$. Note that we only use the tuples from the gold standard where the NLP-based system may hope to make a prediction (i.e. people in the tuple communicate via email). This system, NLP_{Gilbert} achieves an accuracy of 82.37% compared to the social network-based approach (SNA_T) which achieves a higher accuracy of 87.58% on the same test set T. This comparison shows that SNA-based approach out-performs the NLP-based approach even if we evaluate on a much smaller part of the gold standard, namely the part where an NLP-based approach does not suffer from having to make a random prediction for nodes that do not comunicate via email.

System	Test set	# test points	%Acc
UB_{NLP}	G	13,724	59.61
NLP_{Gilbert}	T	2604	82.37
SNA_T	T	2604	87.58
SNA_G	G	13,724	83.88

Table 2: Results of four systems, essentially comparing performance of purely NLP-based systems with simple SNA-based systems.

6 Future Work

One key challenge of the problem of predicting domination relations of Enron employees based on their emails is that the underlying network is incomplete. We hypothesize that SNA-based approaches are sensitive to the *goodness* with which the underlying network represents the true social network. Part of the missing network may be recoverable by analyzing the content of emails. Using sophisticated NLP techniques, we may be able to *enrich* the network and use standard SNA metrics to predict the dominance relations in the gold standard.

Acknowledgments

We would like to thank three anonymous reviewers for useful comments. This work is supported by NSF grant IIS-0713548. Harnly was at Columbia University while he contributed to the work.

References

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011a. Extracting social power relationships from natural language. In *ACL*, pages 773–782. The Association for Computer Linguistics.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011b. Extracting social power relationships from natural language. *ACL*.

Mooi-Choo Chuah and Alexandra Coman. 2009. Identifying connectors and communities: Understanding their impacts on the performance of a dtn publish/subscribe system. *International Conference on Computational Science and Engineering (CSE '09)*.

Germán Creamer, Ryan Rowe, Shlomo Hershkop, and Salvatore J. Stolfo. 2009. Segmentation and automated social hierarchy detection through email network analysis. In Haizheng Zhang, Myra Spiliopoulou, Bamshad Mobasher, C. Lee Giles, Andrew Mccallum, Olfa Nasraoui, Jaideep Srivastava, and John Yen, editors, *Advances in Web Mining and Web Usage Analysis*, pages 40–58. Springer-Verlag, Berlin, Heidelberg.

Christopher Diehl, Galileo Mark Namata, and Lise Getoor. 2007. Relationship identification for social network discovery. *AAAI '07: Proceedings of the 22nd National Conference on Artificial Intelligence*.

Jana Diesner, Terrill L Frantz, and Kathleen M Carley. 2005. Communication networks from the enron email corpus it's always about the people. enron is no different. *Computational & Mathematical Organization Theory*, 11(3):201–228.

Eric Gilbert. 2012. Phrases that signal workplace hierarchy. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW)*.

Bryan Klimt and Yiming Yang. 2004. Introducing the enron corpus. *In First Conference on Email and Anti-Spam (CEAS)*.

Galileo Mark S. Namata, Jr., Lise Getoor, and Christopher P. Diehl. 2007. Inferring organizational titles in online communication. In *Proceedings of the 2006 conference on Statistical network analysis*, ICML'06, pages 179–181, Berlin, Heidelberg. Springer-Verlag.

Sebastian Palus, Piotr Brodka, and Przemysław Kazienko. 2011. Evaluation of organization structure based on email interactions. *International Journal of Knowledge Society Research*.

Ryan Rowe, German Creamer, Shlomo Hershkop, and Salvatore J Stolfo. 2007. Automated social hierarchy detection through email network analysis. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 109–117.

Jitesh Shetty and Jaffar Adibi. 2004. Ex employee status report. `http://www.isi.edu/~adibi/Enron/Enron_Employee_Status.xls`.

Jen Yuan Yeh and Aaron Harnley. 2006. Email thread reassembly using similarity matching. *In Proceedings of CEAS*.