Analyzing Organizational Structures Using Social Network Analysis

Chuanlei Zhang[1], William B. Hurst[1], Rathinasamy B. Lenin[2], Nurcan Yuruk[1], and Srini Ramaswamy[2]

[1] Department of Applied Science, University of Arkansas at Little Rock, 2801 S. University Avenue, Little Rock, Arkansas 72204, USA
{cxzhang,wbhurst,nxyuruk}@ualr.edu
[2] Department of Computer Science, University of Arkansas at Little Rock, 2801 S. University Avenue, Little Rock, Arkansas 72204, USA
rblenin@ualr.edu, srini@acm.org

Abstract. Technological changes have aided modern companies to gather enormous amounts of data electronically. The availability of electronic data has exploded within the past decade as communication technologies and storage capacities have grown tremendously. The need to analyze this collected data for creating business intelligence and value continues to grow rapidly as more and more apparently unbiased information can be extracted from these data sets. In this paper we focus in particular, on email corpuses, from which a great deal of information can be discerned about organization structure and their unique cultures. We hypothesize that a broad based analysis of information exchanges (ex. emails) among a company's employees could give us deep information about their respective roles within the organization, thereby revealing hidden organizational structures that hold immense intrinsic value. Enron email corpus is used as a case study to predict the unknown status of Enron employees and identify homogeneous groups of employees and hierarchy among them within Enron organization. We achieve this by using classification and cluster techniques. As a part of this work, we have also developed a web-based graphical user interface to work with feature extraction and composition.

Keywords: Business intelligence, organizational hierarchies, classification, clustering, Enron email corpus.

1 Introduction

Technological changes have aided modern companies to gather enormous amounts of data electronically. In this paper we focus in particular, on email corpuses, from which a great deal of information can be discerned about organization structure and their unique cultures. It can also be used as a 'window' by entities such as private equity firms, insurance companies, or banks that aspire to do due diligence in understanding a company's culture, be it in legal, regulatory or other needs. We hypothesize that a broad based analysis of information exchanges (emails in this instance)

A. Albani, J. Barjis, and J.L.G. Dietz (Eds.): CIAO!/EOMAS 2009, LNBIP 34, pp. 143–156, 2009.

among a company's employees could give us deep information about their respective roles within the organization, thereby revealing hidden organizational structures that hold immense intrinsic value. Therefore, these corpuses can be considered as a crucial value-added proposition, which helps one decipher the emerging social structures, as well as potentially identify key personnel (rising stars) within an organization. Such analysis is known as Social Network Analysis (SNA). From a purely business perspective, such type of analysis helps us avoid bias in judging / evaluating the importance of each and every player (employee) of an organization, and providing a detailed view that does not solely depend on key administrative decision makers in modern day hierarchical organizational structures. As a preliminary study, we have earlier used SNA in analyzing software developer roles open-source software (OSS) development to identify key developers and coordinators in making OSS systems more successful [1].

The analysis of social networks has focused on a set of important, but relatively simple measures of network structure [2]; these include issues such as degree distributions, degree correlations, centrality indices, clustering coefficients, subgraph (motif) frequencies, preferential attachment, node duplications, degree distributions and correlations. Recently researchers have begun studying wider community structures in networks and issues such as interconnectedness, empirical relationships, weak community links, collaboration, modularity and community structures [3]. SNA in electronic media essentially involves "Link Mining". Link mining is a set of techniques which is used to model a linked domain using different types of network indicators [4]. A recent survey on link mining can be found in [5]. Its applications include NASDAQ surveillance [6], money laundering [7], crime detection [8], and telephone fraud detection [9]. In [10], the authors showed that customer modeling is a special case of link mining.

The public availability of Enron Corporation's email collection, released during the judicial proceedings against this corporation, provides a real rich dataset for research [11, 12]. In [13, 14], the authors used Natural Language Processing techniques to explore this email data. In [15], the authors used SNA to extract properties of the Enron network and identified the key players during the time of Enron's crisis. In [16], the authors analyzed different hierarchical levels of Enron employees and studied the patterns of communication of the employees among these hierarchies. In [17], the authors used a thread analysis to find out employees' responsiveness. In [18], the authors used an entropy model to identify the most relevant people. In [19], the authors proposed a method for identity resolution in the Enron email dataset. In [20], the authors deployed a cluster ranking algorithm based on the strength of the clusters to this dataset. In [21], the authors provided a novel algorithm for automatically extracting social hierarchy data from electronic communication behavior.

In this paper, we apply SNA to identify different social groups among employees based on the emails they exchanged and attempt to predict organizational structure that emerges from such complex social networks. Such an analysis can be very useful from an economic perspective for reasons such as business strategy, competition, multiplayer perspectives, enabling leadership and innovation support structures, stability and societal benefit. Our work is different from that of most of the earlier works, in that it is significantly more comprehensive. Specifically, it focuses on the following two issues, which are significant value-added propositions for any organization.

1. First, we mine the email corpus to collect the data such as counts of the emails which were exchanged between Enron employees using the To, Cc and Bcc fields, different combinations and ratios of these counts, and the response ratio to emails sent by each employee. The development and use of composite features of these various fields (of email data) in our work is significantly different from all reported earlier work. We use clustering algorithms [22] and classification algorithms [23] to analyze the data in order to identify homogeneous groups of employees and to predict the designation status of employees whose status were undocumented in the Enron database report. For clustering we used matlab and for classification we used weka [24]. Furthermore, we have developed a web-based Graphical User Interface (GUI) that can work with different classifiers to automate feature selection and composition, so that we can interactively carry out the classification analysis. The validity of clusters is demonstrated using different widely used statistical measures [25-28]. The validation of classification-based prediction is done by predicting the designation status of employees, for whom the status is known.
2. Second, we use prediction techniques to identify employees who may be performing roles that are inconsistent with other employees in similar roles within the organization. We hypothesize, that these roles may be associated with either more leadership responsibilities or with more mundane day to day operational responsibilities that keep the organization focused on its core capabilities. Such personnel tend to play either a critically vital role in the organization in helping it accomplish its goals and missions (through leadership, mentoring and training of junior employees) or, are leading indicators of poor performers. This implies that the remaining employees perform within well-defined 'bounded' roles in the organization.

The paper is organized as follows. In section 2, we briefly talk about Enron email corpus and features we extracted from this corpus. In section 3, we discuss about clustering techniques we used and about the GUI we developed to work with weka. In section 4, we discuss the results, and in section 5 we conclude this work and discuss possible future work in this direction.

2 Enron Email Corpus

The Enron email dataset was initially made public by Federal Energy Regulation Commission. The raw dataset can be found at [11]. There are different versions of the datasets processed by different research groups. We collected our dataset from [12], as a MYSQL dump, because the authors cleaned this dataset by removing duplicate emails and processed invalid email address. In this dataset, there are a total of 252759 / 2064442 email messages being sent to / received by a total of 151 Enron employees. Information about each of these employees such as full name, designation, email address, emails sent, emails received, subject and body of emails, and references to these messages is maintained in different tables. For 56 employees, the designation information is unavailable and they are marked as 'N/A' in the dataset. In this paper, we try to establish the designation status of these 56 employees. In Table 1, we tabulate the assigned employee identity numbers, and designation status of 151 employees of Enron [29].

Table 1. Details of Enron employees

ID	Status	ID	Status	ID	Status
1	Director	51	Vice President	101	Director
2	Director	52	Employee	102	N/A
3	Employee	53	CEO	103	Vice President
4	Manager	54	President	104	N/A
5	Employee	55	N/A	105	N/A
6	Employee	56	N/A	106	Director
7	Employee	57	N/A	107	President
8	Employee	58	Trader	108	Manager
9	Employee	59	N/A	109	Manager
10	N/A	60	Trader	110	Managing Director
11	N/A	61	Employee	111	N/A
12	Employee	62	Employee	112	N/A
13	N/A	63	N/A	113	N/A
14	Vice President	64	N/A	114	Employee
15	Employee	65	Managing Director	115	N/A
16	N/A	66	President	116	Employee
17	Manager	67	N/A	117	Manager
18	N/A	68	Vice President	118	Employee
19	Employee	69	Vice President	119	Employee
20	N/A	70	Employee	120	Director
21	Director	71	N/A	121	Director
22	Trader	72	N/A	122	Employee
23	Vice President	73	Employee	123	Manager
24	Manager	74	In House Lawyer	124	Trader
25	Employee	75	N/A	125	Trader
26	Director	76	N/A	126	Vice President
27	Employee	77	Employee	127	CEO
28	N/A	78	Vice President	128	Employee
29	Vice President	79	N/A	129	N/A
30	N/A	80	N/A	130	Director
31	N/A	81	Employee	131	Trader
32	Vice President	82	N/A	132	Trader
33	N/A	83	Vice President	133	Trader
34	N/A	84	N/A	134	Trader
35	Vice President	85	Employee	135	N/A
36	President	86	N/A	136	Employee
37	Vice President	87	Employee	137	Vice President
38	Vice President	88	Trader	138	Trader
39	N/A	89	N/A	139	In House Lawyer
40	Trader	90	Manager	140	Employee
41	N/A	91	N/A	141	N/A
42	Manage	92	Vice President	142	Employee
43	N/A	93	Employee	143	Employee
44	Vice President	94	N/A	144	N/A
45	N/A	95	N/A	145	N/A
46	CEO	96	Vice President	146	N/A
47	Employee	97	N/A	147	In House Lawyer
48	N/A	98	N/A	148	N/A
49	N/A	99	N/A	149	Employee
50	N/A	100	Employee	150	N/A
				151	Employee

The dataset extractions where performed at two different levels of discrimination: 1) Localized Email communications (between Enron employees) and 2) Global Email communications (Email involving Enron employees on a global scale). For the former level, email communications among Enron employees were pulled out of the database. For the latter case, a closer level of discrimination was observed, since an Enron employee could be involved in the Email messages from four different levels of abstraction: To, Cc, Bcc, and From. Scripts (in Perl) were used to connect to the MYSQL database tables and extract the data through table joins to fulfill the data requirements for analysis. Once the data has been extracted out, a secondary set of scripts were used to place the data in formats suitable for the types of analysis planned to be performed. In these final steps the data was placed into tab delimited data files and comma delimited files; with subsequent use either in raw data form, or in matrix type summary formats.

3 Analysis

In this section we discuss briefly the clustering and classification algorithms and features sets we collected from Enron email corpus for these algorithms.

We use classification analysis to predict the designation status of employees whose status are reported as 'N/A' in the email corpus. We use k-means, density based expectation maximization (EM), and tree-random forest techniques for classification analysis [23]. Classification algorithms are based on supervised learning methodology; which assumes the existence of a teacher-fitness function or some other external method of estimating the proposed model. The term "supervised" means "the output values for training samples are known (i.e., provided by a 'teacher')" [30]. In our analysis, we use employee's records for which designation status is known to train the algorithm and use the model obtained from the trained data set to predict the status of employees with 'N/A' values in their designation fields. Validation of prediction using these classification techniques is done using the available designation information of 95 (out of a total of 151) employees, i.e the status of 56 employees was unspecified). We use clustering analysis to identify homogeneous groups, of employees within the Enron organization. To achieve this we use k-means and fuzzy c-means clustering algorithms [22]. Each cluster has a centroid, which is a vector that contains the mean value of each feature for the employees within the cluster. Once clusters are identified, we create a centroid matrix and use hierarchical clustering to identify the similarities and hierarchies among different clusters. It is important to validate cluster results to test whether the right number of clusters is chosen and to test whether cluster shapes correspond to the right groupings in the data.

3.1 Finding the Optimal Number of Clusters

Techniques such as the silhouette measure, partition coefficient, classification entropy, partition index, separation index, and Xie and Beni's index are used to find the optimal number of clusters [25-28]. The silhouette measure is a measure of how close each point in one cluster is to points in the neighboring clusters. The partition coefficient measures the amount of overlap between clusters. Xie and Beni's index quantifies the

ratio of the total variation within clusters and separation of clusters. Partition index is the ratio of the sum of compactness and separation of clusters. The classification entropy measures the fuzziness of the cluster partition. Separation index uses a minimum-distance separation for partition validity. The optimal number of clusters is achieved when the first three measures (silhouette measure, partition coefficient, and Xie and Benn index) attain the first local maxima and the later three measures (partition index, classification entropy, and separation index) attain their first local minima.

It is to be noted that not all these techniques are designed to be used with all the chosen clustering techniques. For instance, the silhouette measure is applicable for all clustering techniques whereas partition coefficient and classification entropy are most suitable for fuzzy c-mean algorithm. However, by applying several of these six validation measures together, we can obtain an optimal number of clusters by comparing and contrasting the choices for the number of clusters that concurrently satisfy a majority of these choices.

3.2 Feature Set Identification

For each identifiable employee in the database, we identified nine specific features that could be used for clustering and classification. These features are tabulated in Table 2 and we use these features in our cluster and classification analysis that is reported in section 4. As indicated earlier, the development and use of the ratio features is significantly different from all reported earlier work on the Enron corpus. The major improvement that the use of ratios accomplish is to remove some of the undesirable 'biasing' effects of the raw data collected.

4 Results

In this section, we provide results of our classification and clustering analysis. The consolidated statistics, based on the designation status given in Table 1, is tabulated in Table 3. For our clustering, analysis and prediction we use the groupings as identified in Table 4.

4.1 Classification Analysis Results

For classification analysis, to predict the status of employees with 'N/A' fields, we first tested the dataset using k-means, density based expectation maximization (EM) and tree-random forest techniques. For the analysis we created six different employee groups based on their designation as shown in Table 4 and further assigned numeric ranks to these groups.

We used the features listed in Table 2 for the employees with known designation status as the training data to train the classification algorithms. Steps involved in the process of classification are explained pictorially in Fig. 1. Different techniques (k-means, density based expectation maximization (EM) and tree-random forest) were used to test accuracy. While both k-means and density based EM did not produce an accuracy rate of more than 60%, the tree-random forest technique, produced best

Table 2. Features extracted for Enron email corpus

Feature Id	Feature Description
f_1	Number of emails sent in Cc field
f_2	Number of responses received for emails mentioned for feature f_1
f_3	Cc Response Ratio: f_1 / f_2
f_4	Number of emails sent in To field
f_5	Number of responses received for emails mentioned for feature f_4
f_6	To Response Ratio: f_4 / f_5
f_7	Number of emails received in the To field
f_8	Number of emails received in the Cc field
f_9	Informational Response Ratio: f_7 / f_8

Table 3. Consolidated statistics of Enron employees based on their status

Status	Count
N/A	56
CEO	3
Director	9
Employee	35
In house lawyer	3
Manager	9
Managing Director	2
President	4
Vice President	18
Trader	12

Table 4. Six groups of employees based on their designation

Status	Group Name	Rank
Director, Manager	Middle-Management	3
Vice President, CEO, President	Upper-Management	1
Employee	Employee	4
In House lawyer	In House lawyer	4
Trader	Trader	4
Managing Director	Managing Middle-Management	2

accuracy of 100%. Given the relatively limited size of the data this is expected and hence, the tree-random forest technique is used to predict the 'N/A' status of employees.

Some interesting observations are to be made in the results of predicting employee ranks that were 'N/A'. These are shown in Table 5. In column 2, we predicted the

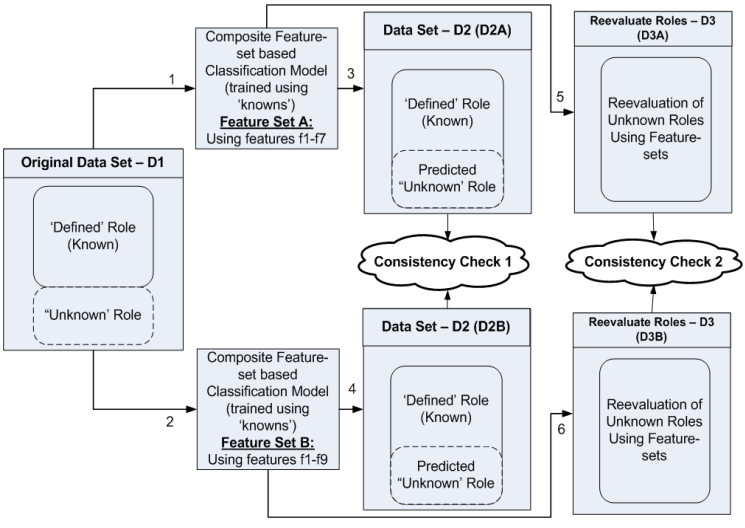


Fig. 1. Steps in classification analysis

rank of 'N/A' employees without using the ratio features f3, f6, and f9 (from table 2). In column 3 of Table 5, the results of predicting the rank of 'N/A' employees without using two key 'differentiation' features – i.e. the feature f8 - where the recipient were sent a 'cc' and the ratio f9, where the ratio of emails sent is computed. In column 4, the prediction of ranks of these 'N/A' employees using all features is presented, while in column 5, we present the results of predicting the roles of all employees – including ranks of those that were previously specified. The objective was to identify and predict some of the subtle differences in the internal roles played within the organizational by these large number of employees that were reported as 'N/A'. From column 5, using the most complete set of features it can be seen that about 41 of these play a upper (23) or middle (18) management ranked employees. Only 15 played the role of an employee (12) or trader (3). This is in stark contrast to not using ratios as features, where 32 of them ranked as either employees (27) or traders (5) in comparison to 23 that ranked as middle (8) or upper management (15) designations.

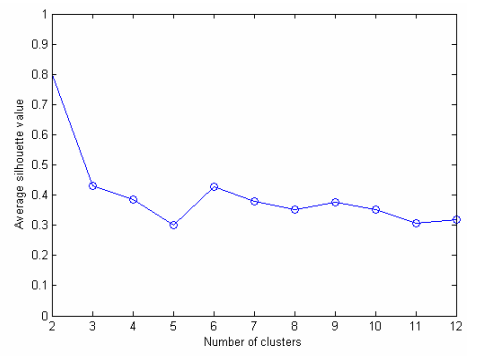
4.2 Cluster Analysis Results

In this section, we apply cluster analysis for the email dataset to identify homogeneous groups, of employees within the Enron organization. To achieve this we use fuzzy c-means clustering algorithms. We use the silhouette measure, partition coefficient, classification entropy, and Xie and Beni's index, to validate the right number of clusters and fuzzyness of the cluster partition. Once the clusters are identified, we create a centroid matrix and use hierarchical clustering to identify hierarchies among different clusters.

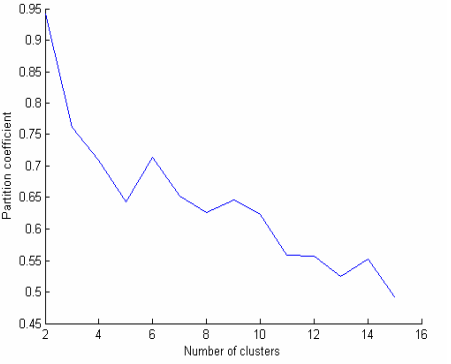
Table 5. Predicting Roles of 'N/A' Designations

			All Features	
Emp ID	Without Ratio	Seven Features	1st round prediction	2nd round prediction
10	Employee	Employee	Upper-Management	Upper-Management
11	Employee	Employee	In-House-Lawyer	Employee
13	Employee	Employee	Upper-Management	Upper-Management
16	Employee	Employee	Upper-Management	Upper-Management
18	Employee	Employee	Upper-Management	Upper-Management
20	Employee	Employee	Employee	Employee
28	Employee	Employee	Middle-Management	Upper-Management
30	Employee	Employee	Upper-Management	Trader
31	Employee	Employee	Upper-Management	Middle-Management
33	Employee	Employee	Employee	Upper-Management
34	Employee	Employee	Managing-Middle-Management	Middle-Management
39	Employee	Employee	Upper-Management	Upper-Management
41	Employee	Employee	Upper-Management	Middle-Management
43	Employee	Employee	Middle-Management	Middle-Management
45	Employee	Employee	Trader	Middle-Management
48	Employee	Employee	Upper-Management	Upper-Management
49	Employee	Employee	In-House-Lawyer	Upper-Management
50	Employee	Employee	Middle-Management	Middle-Management
55	Employee	Employee	Upper-Management	Upper-Management
56	Employee	Employee	Managing-Middle-Management	Middle-Management
57	Employee	Employee	Employee	Employee
59	Employee	Employee	Employee	Employee
63	Employee	In-House-Lawyer	Upper-Management	Upper-Management
64	Employee	Managing-Middle-Management	Trader	Middle-Management
67	Employee	Middle-Management	Upper-Management	Upper-Management
71	Employee	Middle-Management	Middle-Management	Trader
72	Employee	Middle-Management	Employee	Trader
75	In-House-Lawyer	Middle-Management	Upper-Management	Upper-Management
76	Middle-Management	Middle-Management	In-House-Lawyer	Upper-Management
79	Middle-Management	Middle-Management	Employee	Middle-Management
80	Middle-Management	Middle-Management	Upper-Management	Upper-Management
82	Middle-Management	Middle-Management	Managing-Middle-Management	Employee
84	Middle-Management	Trader	Upper-Management	Upper-Management
86	Middle-Management	Trader	Employee	Employee
89	Middle-Management	Trader	Upper-Management	Upper-Management
91	Middle-Management	Upper-Management	Middle-Management	Upper-Management
94	Trader	Upper-Management	Trader	Middle-Management
95	Trader	Upper-Management	Employee	Upper-Management
97	Trader	Upper-Management	Middle-Management	Middle-Management
98	Trader	Upper-Management	Managing-Middle-Management	Upper-Management
99	Trader	Upper-Management	Trader	Employee
102	Upper-Management	Upper-Management	In-House-Lawyer	Middle-Management
104	Upper-Management	Upper-Management	Upper-Management	Middle-Management
105	Upper-Management	Upper-Management	Trader	Employee
111	Upper-Management	Upper-Management	Employee	Upper-Management
112	Upper-Management	Upper-Management	Middle-Management	Middle-Management
113	Upper-Management	Upper-Management	Middle-Management	Upper-Management
115	Upper-Management	Upper-Management	Employee	Employee
129	Upper-Management	Upper-Management	Managing-Middle-Management	Middle-Management
135	Upper-Management	Upper-Management	Upper-Management	Upper-Management
141	Upper-Management	Upper-Management	Upper-Management	Employee
144	Upper-Management	Upper-Management	Employee	Employee
145	Upper-Management	Upper-Management	Middle-Management	Middle-Management
146	Upper-Management	Upper-Management	Middle-Management	Middle-Management
148	Upper-Management	Upper-Management	Middle-Management	Middle-Management
150	Upper-Management	Upper-Management	Employee	Employee

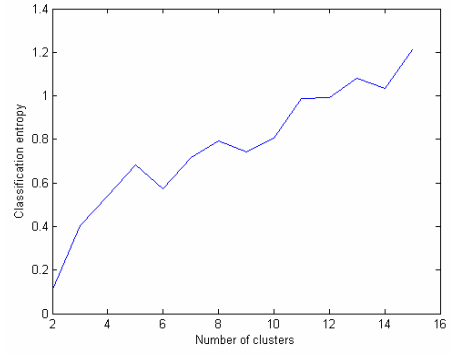
Since the dataset is limited and the expected number of clusters is not too large, we use the fuzzy c-means clustering algorithm to identify the homogeneous group of employees using the features given in Table 2. Based on the validity measures shown in Fig. 2, the optimal cluster size is found to be 6, and these clusters are displayed in Fig. 3. For these 6 clusters, the hierarchy tree is shown in Fig. 4. Clusters 2 and 6 are in the same (bottommost) level, cluster 4 is in the next level, clusters 1 and 5 are in the next level, and cluster 3 is in the top level. The tree structure can be interpreted as follows based on the number of emails exchanged among them: Members of clusters 2 and 6 possess similar characteristics and report to members of cluster 4 who in turn



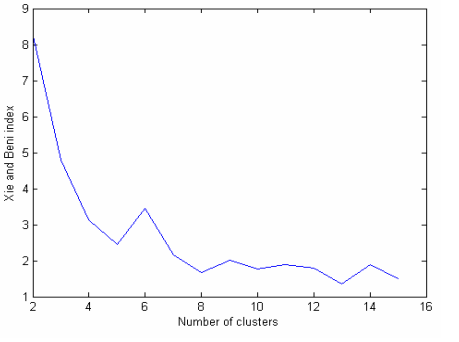
(a) Average silhouette values



(b) Partition coefficient measures



(c) Classification entropy measure



(d) Xie and Benn index

Fig. 2. Validation for optimal number of clusters for fuzzy c-means clustering technique

report to members of cluster 3. Strangely though clusters 1 and 5 possessed similar characteristics, they stay independent of other clusters.

The topmost cluster in the tree is 3 and whose members' ids and status are tabulated in Table 6. For the predicted status, the status group based on Table 4 is provided. Clusters 1 and 5 are isolated from other four clusters. The members of cluster 1 are 75 and 107, and members of cluster 5 are 48, 67, 69, and 73. Except 73 and 75 (employee and middle management), all other members belong to upper management group. We know that member 107 is the President and its interesting that grouped with the president is employee 75 (originally N/A). This person (75) can be considered as a key connection point for other clusters. It is also clear that the management groups are isolated from other groups (clusters 2, 3, 4 and 6) which may be an ideal case for any organization. Similarly, the employee member 73 belongs to cluster 5; which is dominated by upper-management members and function closely with President's group and hence he/she can be considered to be a key connection between cluster 5 and other clusters.

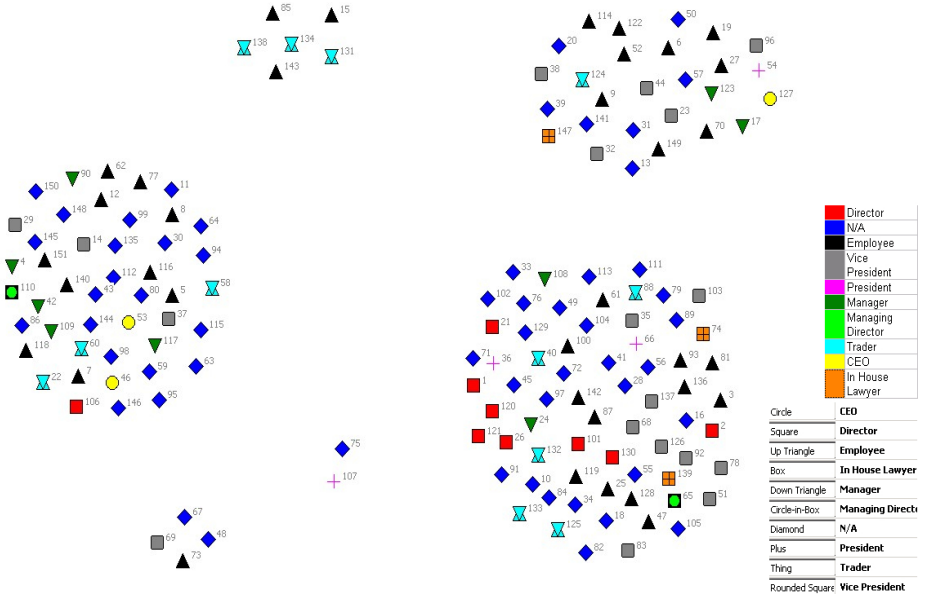


Fig. 3. Identification of employee clusters

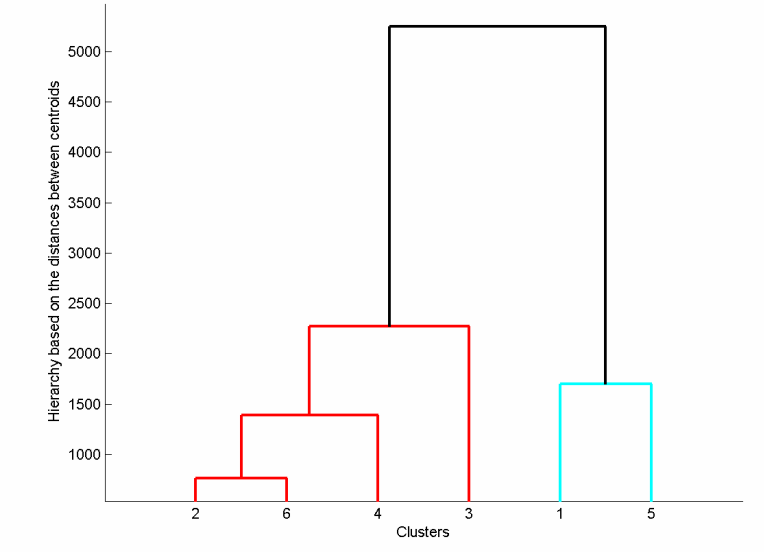


Fig. 4. Hierarchy among clusters (using fuzzy c-means clustering)

The members of cluster 3 (top left in Fig.2) mostly consists of traders and employees. Members in clusters 2, 4 and 6 (top right, bottom right and middle left in Fig. 2) are mixed with members of well-mixed status and hence nothing much can be inferred from this, although members of 2 and 6 work in the same hierarchy level (Fig. 3).

Table 6. Members of top level cluster in the hierarchical tree based on fuzzy c-means clustering technique

ID	Status	Predicted Status (Table 5)
18	N/A	N/A → Upper Management
23	Vice President	*
48	N/A	N/A → Upper Management
67	N/A	N/A → Upper Management
69	Vice President	*
73	Employee	*
75	N/A	N/A → Upper Management
107	President	*
114	Employee	*
137	Vice President	*
147	In House Lawyer	*

Since most of the upper management members are isolated well in clusters 1 and 5 and seem to have acted independent of the other clusters, we can safely conclude that SNA based on the features we extracted from the email corpus verifies our hypothesis: "A broad based analysis of information exchanges (emails in this instance) among a company's employees could give us deep information about their respective roles within the organization, thereby revealing hidden organizational structures that hold immense intrinsic value." In case of Enron, this indicates that the regular employees were probably quite unaware of the emerging problems within the organization. Furthermore, this email corpus also indicates that the presence of management personnel (upper/middle management) personnel who wielded quite some influence, yet performed several 'undefined' roles (N/A's) within the organization.

5 Conclusion

In this paper we carried out a case study of social network analysis on Enron email corpus. Out of 151 employees, there were 56 employees whose status was not reported in the Enron email corpus. As a first step in our analysis, we extracted 9 features from the email corpus. We used these features to predict the unknown status of the employees using the tree random forest classification algorithm from Weka. We further predicted how consistent these 51 employees with respect to their designation status. After predicting the unknown status of employees, we identified homogeneous groups of employees and hierarchy among them using the fuzzy c-mean clustering technique. The results of clustering technique supported our hypothesis that a suitable SNA on electronic data would reveal enough information about strengths and weaknesses of organizations and identify potential employees who played crucial roles in the organization.

In later work, we plan to extend the capability of the web-based GUI that currently leads us perform custom feature composition for analysis, to include clustering / classification analysis using alternate techniques. We plan to identify weighted features based on the response time –ex. for feature such as f2, f4; by modeling response time as a power-law distribution so as to assess relative importance of messages. As appropriate, we plan to use global email communications (emails which have been

exchanged between Enron employees and known outsiders – for example lawyers / bankers) to identify the roles played by outsiders. We plan to extend such analysis to study Linux email corpuses to identify key players in the development of Linux, an effort to identify the success behind this open source software system. Such effort is aimed at mimicking identified successes to understand and replicate it for more effective and efficient software development processes.

Acknowledgments

This work is based in part, upon research supported by the National Science Foundation under Grant Nos. CNS-0619069, EPS-0701890 and OISE 0650939 and Acxiom Corporation under contract #281539. Any opinions, findings, and conclusions or recommendations expressed in this material are those of author(s) and do not necessarily reflect the views of the National Science Foundation of Acxiom Corporation.

References

1. Yu, L., Ramaswamy, S., Zhang, C.: Mining email archives and simulating the dynamics of open-source project developer networks. In: Fourth International Workshop on Enterprise and Organizational Modeling and Simulation, Montpellier, France, pp. 17–31 (2008)
2. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press, Cambridge (1994)
3. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (2008)
4. Senator, T.E.: Link mining applications: Progress and challenges. SIGKDD Explorations 7(2), 76–83 (2005)
5. Getoor, L., Diehl, C.P.: Link mining: A survey. SIGKDD Explorations 7(2), 3–12 (2005)
6. Goldberg, H.G., Kirkland, J.D., Lee, D., Shyr, P., Thakker, D.: The NASD securities observation, news analysis and regulation system (sonar). In: IAAI 2003, pp. 11–18 (2003)
7. Kirkland, J.D., Senator, T.E., Hayden, J.J., Dybala, T., Goldberg, H.G., Shyr, P.: The nasd regulation advanced detection systems (ads). AI Magazine 20(1), 55–67 (1999)
8. Sparrow, M.: The application of network analysis to criminal intelligence: an assessment of the prospects. Social Networks 13, 251–274 (1991)
9. Provost, F., Fawcett, T.: Activity monitoring: noticing interesting changes in behavior. In: Fifth ACM SIGKDD International conference on knowledge discovery and data mining (KDD 1999), pp. 53–62 (1999)
10. Huang, Z., Perlinch, C.: Relational learning for customer relationship management. In: International Workshop on Customer Relationship Management: Data Mining Meets Marketing (2005)
11. Enron, Enron Email Dataset, `http://www.cs.cmu.edu/~enron/`
12. Adibi, J., Shetty, J.: The Enron email dataset database schema and brief statistical report, Information Sciences Institute (2004)
13. Yang, Y., Klimt, B.: The enron corpus: A new dataset for email classification research. In: European Conference on Machine Learning, Pisa, Italy (2004)

14. McCallum, A., Corrada-Emmanuel, A., Wang, X.: The author-recipient-topic model for topic and role discovery in social networks: Experiments with entron and academic email. In: NIPS 2004 Workshop on Structured Data and Representations in Probabilistic Models for Categorization, Whister, B.C. (2004)
15. Carley, K.M., Diesner, J.: Exploration of communication networks from the enron email corpus. In: Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA (2005)
16. Diesner, J., Frantz, T.L., Carley, K.M.: Communication networks from the Enron email corpus. Journal of Computational and Mathematical Organization Theory 11, 201–228 (2005)
17. Varshney, V., Deepak, D.G.: Analysis of Enron email threads and quantification of employee responsiveness. In: Workshop on International Joint Conference on Artificial Intelligence, Hyderabad, India (2007)
18. Adibi, J., Shetty, J.: Discovering important nodes through graph entropy: the case of Enron email database. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Ilinois, U.S.A. (2005)
19. Oard, D.W., Elsayed, T.: Modeling identity in archival collections of email: a preliminary study. In: Third Conference on Email and Anti-spam (CEAS), Mountain View, CA (2006)
20. Bar-Yossef, Z., Guy, I., Lempel, R., Maarek, Y.S., Soroka, V.: Cluster ranking with an application to mining mailbox networks. In: ICDM 2006: Proceedings of the Sixth International Conference on Data Mining, Washington, DC. U.S.A, pp. 63–74 (2006)
21. Rowe, R., Creamer, G., Hershkop, S., Stolfo, S.J.: Automated social hierarchy detection through email network analysis. In: Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007, San Jose, California, USA, pp. 1–9 (2007)
22. Everitt, B.S., Landau, S., Leese, M.: Cluster Analysis, 4th edn. A Hodder Arnold Publication (2001)
23. Izenman, A.J.: Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, 1st edn. Springer, Berlin (2008)
24. Weka. Weka: Data Mining Software in Java, `http://www.cs.waikato.ac.nz/ml/weka/`
25. Bensaid, A.M., Hall, L.O., Bezdek, J.C., et al.: Validity-guided (Re)Clustering with applications to image segmentation. IEEE Transactions on Fuzzy Systems 4, 112–123 (1996)
26. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum press (1981)
27. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, Chichester (1990)
28. Xie, X.L., Beni, G.A.: Validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 3(8), 841–846 (1991)
29. Enron Dataset, `http://www.isi.edu/~adibi/Enron/Enron.htm`
30. Kantardzic, M.: Data Mining: Concepts, Models, Methods, and Algorithms, 1st edn. Wiley/ IEEE (2002)