Classifying Organizational Roles using Email Social Networks

Abtin Zohrabi, Fatemeh Razzaghi, and Ali A. Ghorbani

Abstract

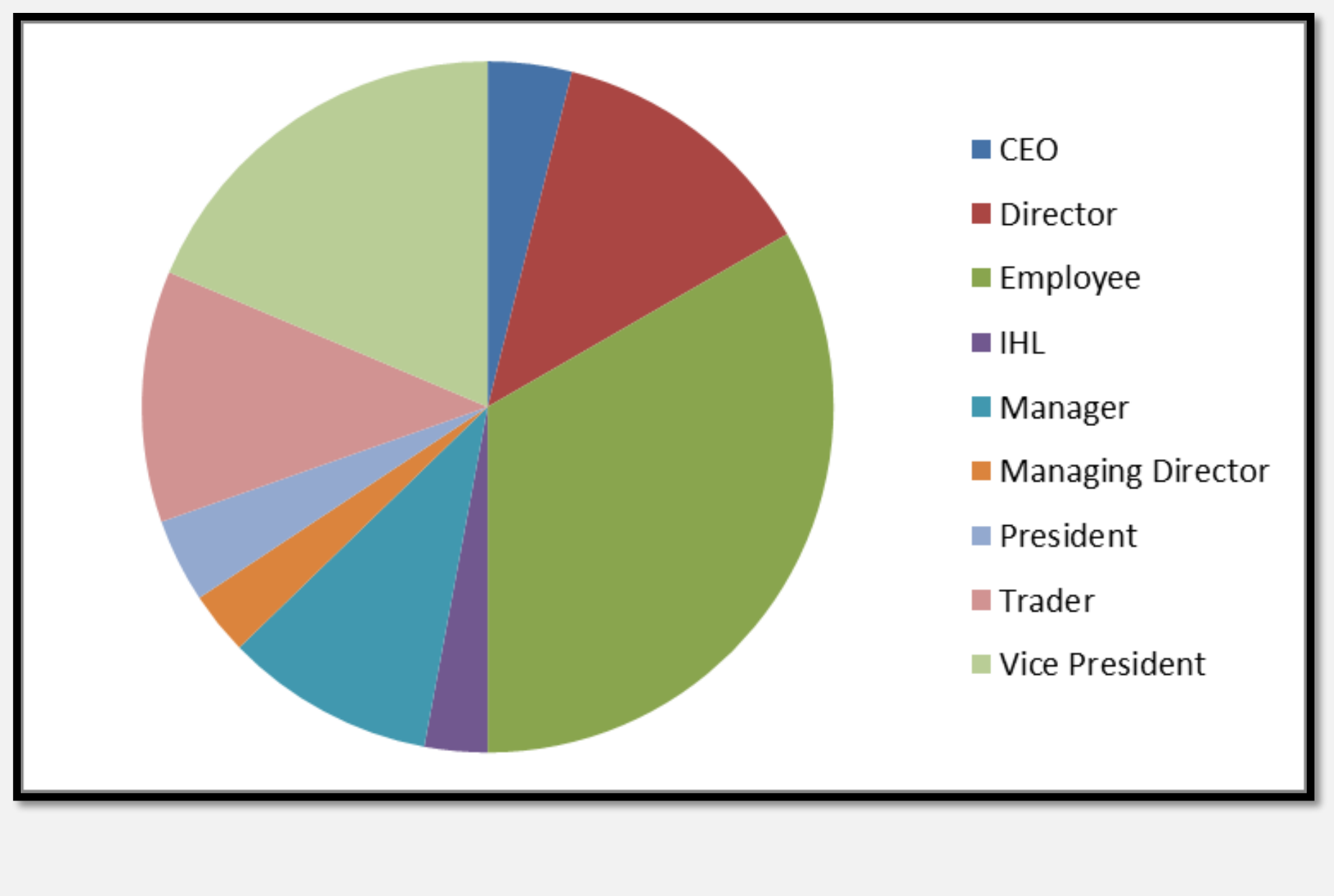
This work addresses the problem of role classification, which is related to classifying and grouping email users into a collection of organizational roles. This classification can be used in designing modern email clients by adding an Inbox prioritizing feature that can predict the role of a sender to the recipient of an email. A comprehensive study has been done on the social network of the Enron dataset. For classifying organizational roles, a feature vector containing a set of social network metrics and interaction-based features reflecting users' engagingness and
responsiveness in their community is created. In turn, a Neural Network classifier has been built based on the extracted features for classifying organizational roles that resulted in 63:57% of accuracy.

Introduction

Role classification is defined as finding the role of a user in an organizational setting from a communication network (i.e. email). Email clients can take advantage of this technique in order to prioritize the incoming emails based on the role of the email's sender within the specific organizational setting.
A series of questions arise: Can we extract some numerical measures out of an organization's email interactions that can contribute to classifying roles of the users? Are these features meaningful enough for accurate discovery and prediction of roles using only small amounts of labeled training data and a limited number of interaction traces?. This work has tackled these fundamental problems by an extensive study on the Enron email dataset.
Most previous research that are conducted on the Enron email corpus have focused on Natural Language Processing of the data for classification of the emails, dataset mapping of Enron's users, and quantitative analysis inside the Enron dataset.

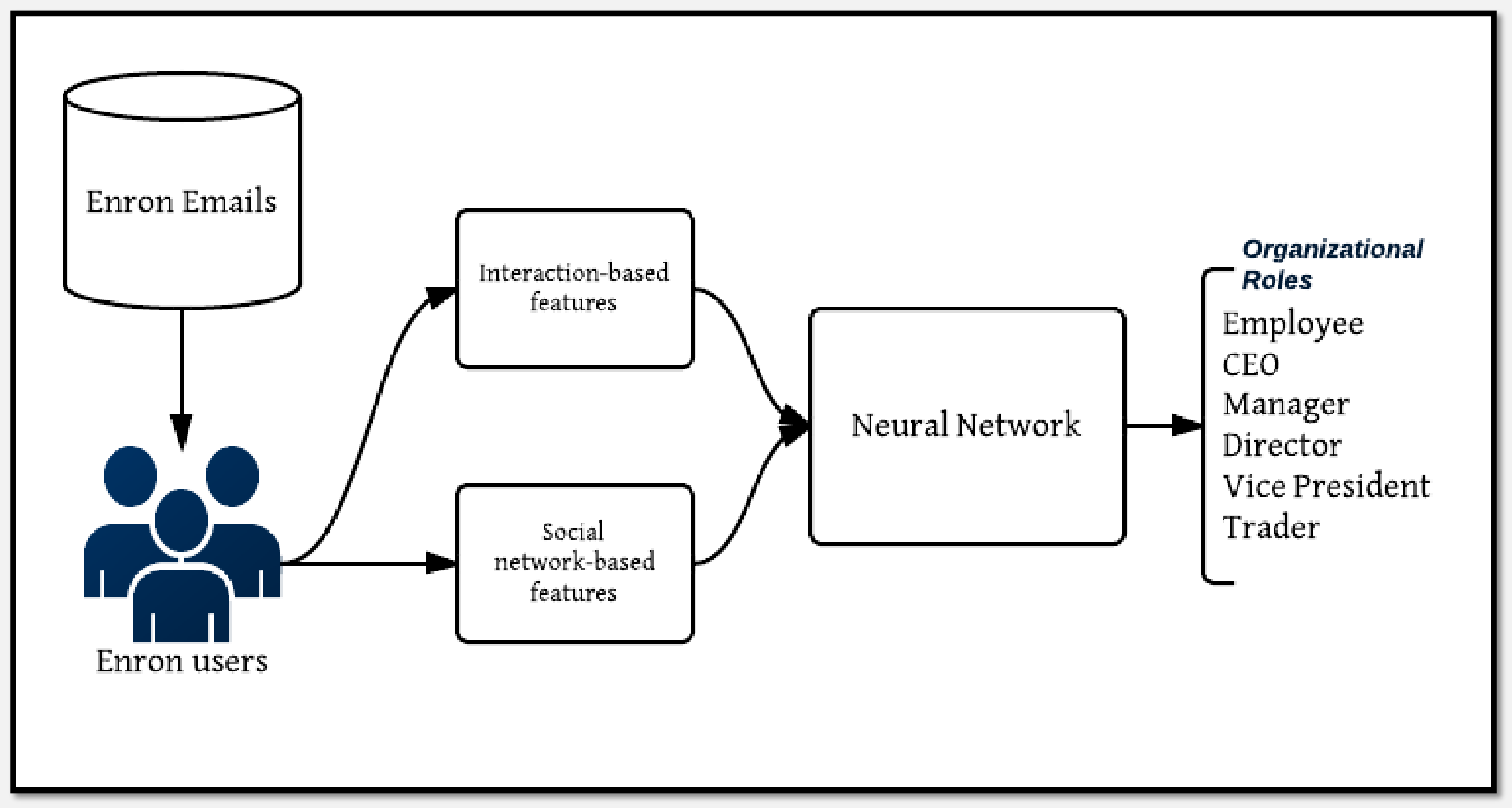
Dataset

The only publicly available dataset which has been used in email research is the Enron corpus. To our knowledge there exists only one list1 that indicates the roles of Enron users: 103 out of 149 users' email addresses and the organizational roles are specified. For the preprocessing, the names and email addresses were examined and found that 60 of selected users had second email addresses. We extracted all of messages related to these addresses and unified the whole dataset.



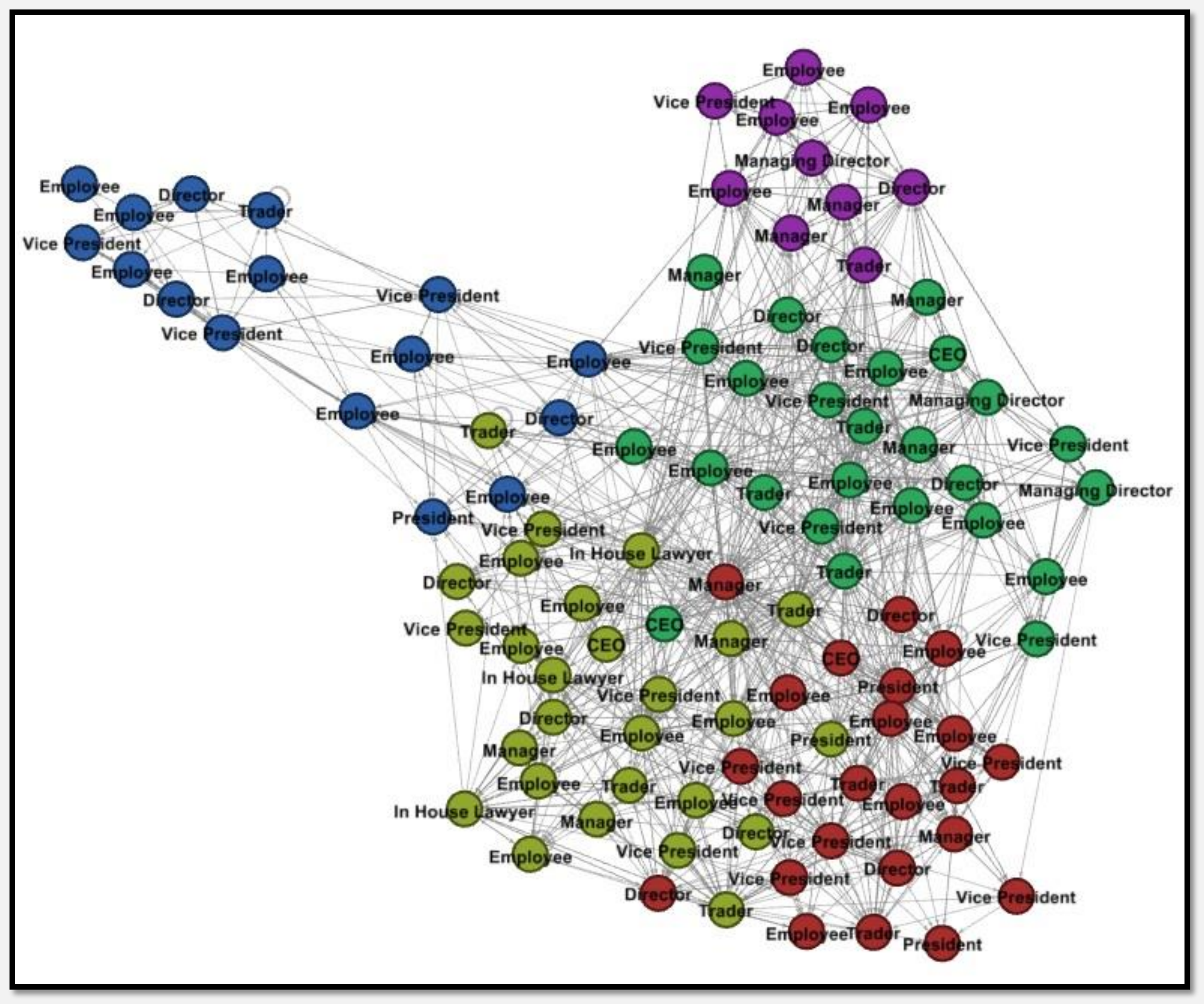
How to Classify

Given a user, described by a set of features, the goal is to build a predictive model which can classify a user's role in an organization. Through a series of experiments, first the social network of the Enron users is analyzed for selecting a set of best feature measures. Then, using these selected features a model is built for organizational role classification.



Social Network Analysis on Enron Dataset

A comprehensive analysis has been done to select the most informative social network features. Social communities are discovered and role communications are studied in this context. Moreover to evaluate if these communities are meaningful and distinguishable in the Enron organizational setting, a topic discovery is made to evaluate these communities.



The Email social network considered as a graph where vertices represent the Enron users and their related roles and edges shows the existence of sent(or received) messages between two users. Applying the Newman clustering algorithm the whole network divided into five different communities.
To evaluate the quality of the communities, emails in each community have been analyzed in isolation in order to find the most frequently discussed terms in that community.

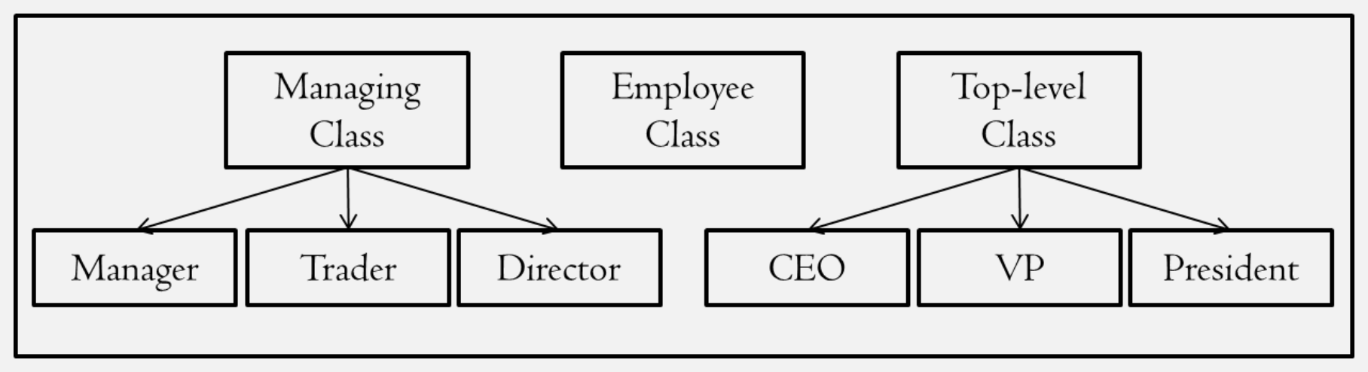
Feature set

Two sets of features have been used for the purpose of Organizational Role Classification:

(1) In-degree	$\frac{1}{\lvert C\rvert}\sum_{j=1}^{i} R_{ji}$,where $R_{ji}\in\{0,1\}$, and $\lvert C\rvert$ is the total number of contacts.
(2) Out-degree	$\frac{1}{\lvert C\rvert}\sum_{j=1}^{i} R_{ij}$
(3) Total-degree	$\frac{1}{\lvert C\rvert}\sum_{j=1}^{\lvert C\rvert}\lceil\frac{R_{ij}+R_{ji}}{2}\rceil$
(4) Clustering Coefficient	$\frac{1}{v}\sum_{i\in Nbr(v)}\sum_{j\in Nbr(n),j\neq i} R_{ij}$
(5) Betweenness	$\frac{1}{(v-1)(v-2)}\sum_{j=1,j\neq i}^{\lvert n\rvert}\sum_{k=1,k\neq j,k\neq i}^{\lvert n\rvert}\frac{\sigma_{jk}(i)}{\sigma_{jk}}$, where σ_{jk} is the number of shortest path between j and k that goes through i[15].
(6) Email Reply-Sent count	$\frac{\lvert RepT(u_i)\rvert}{\lvert Sent(u_i)\rvert}$, where $RepT(u_i)$ are the emails replied to u_i earlier emails and $Sent(u_i)$ is the total number of emails sent by this user.
(7) Email Reply-Received count	$\frac{\lvert RepB(u_i)\rvert}{\lvert Recieved(u_i)\rvert}$, where $RepB(u_i)$ are the emails replied by u_i and $Received(u_i)$ is the total number of emails received by this user.
(8) HITS Authority	Measures the global ratings of a contact inside the whole network

Experiments and Results

For evaluating the selected feature set, EM algorithm as a cluster analysis method is applied which finds the best possible number of clusters through an iterative process.
To overcome the problem of small sample size and considering the issue of skewed role categories we have decided to merge the categories based on their closeness in any organizational setting.



The Neural Network classifier has been trained and the overall accuracy among all the classification is 63.57%.

	TP	FP	Precision	Recall	F-Measure	ROC
Managing Class	0.745	0.221	0.618	0.732	0.675	0.864
Employee Class	0.584	0.124	0.67	0.574	0.608	0.791
Top-level Class	0.568	0.095	0.642	0.55	0.596	0.895
Weighted Avg.	0.636	0.159	0.645	0.629	0.635	0.853
Correctly Classified 63.57%						

This shows that social network measures along with some Interaction-based features can make a strong contribution to classifying roles.

Future Directions

❑ For future work it will be interesting to analyze these communities in time spaces: one can study users and their corresponding feature values in time slices.

❑ To overcome the problem of "small sample size" one can try semi-supervised classifiers like co-training.