Analyzing Interaction Communication Networks in Enterprises and Identifying Hierarchies

Yi Wang*, Marios Iliofotou*, Michalis Faloutsos*, and Bin Wu†
*Department of Computer Science and Engineering
University of California, Riverside, Riverside, CA 92507
Email: {wangyi, marios, michalis}@cs.ucr.edu
†Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia
Beijing University of Posts and Telecommunications, Beijing, China 100876
Email: wubin@bupt.edu.cn

***Abstract*—With the proliferation of electronic modes of communication (e.g., e-mails, short messages), a group of people in an enterprise can form several distinct Communication Interaction Networks, or CINs for short. A CIN is essentially a graph representation of ”who talks to whom” among a group of individuals. In this paper, we conduct an empirical study of two modern enterprises and focus on three main questions: (Q1) how CINs from the two enterprises look, (Q2) how employees use the different available communication modes within an enterprise, and (Q3) how much information we can extract regarding the roles of their participants. We address these questions using empirical CINs from the Enron Corporation and a communication provider, using information from the exchange of e-mails, phone-calls, and short messages (SMS). For Q1, we reveal the following key structural properties that are shared by all the CINs in our study: they have high edge density, high clustering coefficient, and close to zero assortativity coefficient. For Q2, we observe that employees have differences in how they use the various communication modes. This suggests that different CINs capture different behavioral properties within an enterprise. For Q3, we propose HumanRank, a method of ranking individuals based on their importance (e.g., CEOs having higher rank than ordinary employees) using only the interactions between them. Next, using HumanRank, we introduce an unsupervised and parameter-free algorithm that identifies hierarchies by separating managers from ordinary employees. Our algorithm achieves above 70% accuracy and outperforms the state-of-the-art [1].**

I. Introduction

Today’s proliferation of electronic modes of communication (e.g., e-email exchange), increased the ways people interact in an enterprise environment. To facilitate the study of interactions in an enterprise, we represent a group of communicating individuals as a network (or graph), where nodes denote employees and edges capture particular communication, such as a phone-call, between two nodes. We refer to these networks as Communication Interaction Networks, or CINs. With the ever increasing availability of CINs, it is important for people to effectively analyze and understand these data and exploit any information that may be inferred.

In this paper, we focus on three main questions:

- Q1: How do the CINs from two modern enterprises look and which are their common characteristics?
- Q2: How are the different communication modes used within an enterprise? Are there differences between ordinary employees and managers in how they use different communication modes?
- Q3: How much information can we extract regarding the roles of their participants and how can we identify hosts in different levels of the hierarchy (e.g., managers versus ordinary employees)?

We address the above questions using empirical CINs from the Enron Corporation and a communication provider. To the best of our knowledge, we are the first to present similarities between CINs from different enterprises as well as to compare CINs created by different communication modes (short messages and phone calls) within the same enterprise. In the remaining of the paper, all our observations reflect behaviors seen in the two enterprises for which we have data. In the future, we plan to include data from more enterprises in order to identify key properties that capture intrinsic behaviors in modern corporations.

With respect to Q1, we observe that CINs from our two different enterprises have common graph structural features: they have high edge density, high clustering coefficient, and close to zero assortativity coefficient. These similarities are present even though the CINs we study are formed using various communication modes, such as phone call, text messages, and emails. In order to further highlight these features, we generate synthetic random graphs with the same node and edge numbers as our CINs. This allows us to identify key structural properties of our CINs that are not observed in random graphs, even when they exactly share the same degree distribution.

With respect to Q2, we observe that edges in CINs formed by different communication modes, even within the same enterprise can be different. In fact, we observed that only 41% of communicating node pairs use both short messages and phone calls when interacting with each other. Moreover, we noticed that employees with different roles in the enterprise have distinct communication behaviors: for example managers are more likely to utilize multiple modes to contact others whereas ordinary employees use just one.

With regards to Q3, we focus on a very specific question: *Can we distinguish between ordinary employees and managers by using only the graph topology?* Towards this end, we propose **HumanRank** (§V-A), a method of ranking individuals based on their position in the hierarchy (e.g., CEOs having higher rank than ordinary employees). Next, using our HumanRank method, we introduce an unsupervised and parameter-free algorithm that separates managers from ordinary employees. Our algorithm achieves above 70% accuracy in labeling managers and employees. Especially, our method is better in capturing managers and outperforms the state-of-the-art [1] by more than 15%.

The paper is organized as follows. In §II, we review related work and discuss how we differ from it. In §III, we describe the data sets and the metrics we used for analyzing CINs. In §IV, we present the analysis of our graphs, and compare topology features and communication behaviors of same nodes. In §V, we present HumanRank and compare it with others. We conclude the paper in §VI.

II. Related Work

Analyzing complex networks. The study of complex networks has attracted significant interest in the past years. In this paper, we use analysis methodologies similar to others that have been studied: scientific collaboration networks [2], actor collaboration networks [3], online social networks [4][5], the Internet AS-level topology [6], the Web graph [3], and biological networks [7]. Despite the distinct nature of these networks, scientists use similar metrics to capture their structural characteristics. To analyze CINs, we use a combination of metrics proposed in different work. These metrics involve degree and degree correlations [8][9], clustering coefficient [10], assortativity [11], path length, betweenness [12], and network motifs [13].

Ranking nodes and identifying roles in networks. Significant information can be extracted from real-world complex networks. For example, Google employes a PageRank algorithm to assign importance values to web-pages using information from the Web graph [14]. Agarwal et. al [15] presents a framework for ranking network entities by combining connectivity and other source of information, when they are available. Others propose methods for identifying particular roles, but they focus on a different problem. For example, in [16][17] the authors introduce methods for identifying hubs and authorities in networks. Email exchanges in a large enterprise is the topic in [18], but the work is focused on clustering similar users. Moreover, in [18] they do not focus on analyzing the structure of the graph and they do not aim to detect hierarchies, which makes it different from our work.

Inferring hierarchy in enterprise CINs. Next we describe the work that are more related to ours. The release of the Enron email data set in 2004 opens the way for research on ranking employees and inferring their roles (or job descriptions) based on email communication patterns [1]. In addition, they propose a entropy-based ranking method for employees, which is used for comparsion in our paper. Rowe et al. [19] propose a ranking method for employees in the Enron data set, which combined several topological graph features for each node (e.g., its degree). Their method asks users to manually assign weights to features, which makes it hard to use in practice. We applied the method in [19] to our data and used the same weight for all metrics and observed poor ranking results, so we do not include detailed results here for brevity.

III. Datasets and Graph Metrics

A. Datasets

For facilitating the analysis of communication patterns, we represent a group of people as a network $G(V, E)$ (or graph), where nodes (V) represent individuals and edges (E) capture particular interactions. Here, we study enterprise CINs, where nodes are employees in the same enterprise and edges are formed from the exchange of phone-call, short message, or emails among them. We will refer to these CINs as CALL, SMS and EMAIL networks, respectively.

CALL and SMS: These data capture communications among 235 employees from a city branch of a large corporation. For each employee, we have their job titles: manager and ordinary employee. We use this information as ground truth in evaluating our algorithms. The data span over half a year. The SMS data are from the exchange of short messages from the employees' cell-phones. The CALL CIN captures all the phone-call between the same set of employees. For CALL CIN, we have callee and callee and thus we produce directed edges. Unfortunately, SMS CIN is undirected because sender and receiver are not able to be distinguished in the dataset. Edge weight is defined as the sum of communication times between two parcipants during the data span.

EMAIL: This network is extracted from the Enron public email data set, collected by J. Shetty [1]. The public data is comprised of email exchanges among 151 employees that span over 2.5 years. For 101 employees, we have his/her name job title (e.g., CEO, manager) and job description. The job title is unknown for 50 Enron employees. We always include these nodes in our graphs, but we do not use them when we evaluate our methods. Like CALL, EMAIL is directed. However, in order to study CALL, SMS and EMAIL in a uniform way, we consider them as undirected networks except in motif discovery.

B. Graph metrics

In this paper, we focus on following graph metrics: node degree, assortativity [11], node distance, diameter [20], [4], clustering coefficient [10], betweennes [12] and maximal cliques [21]. They are widely used to capture network features in other complex network analysis, so their definitions are not introduced here. In our paper, we use the following symbols to represent them: $\bar{k}$ and $\bar{C}$ are average degree and clustering coefficient. r is asssortativity.

C. Generating synthetic random networks for comparison

In order to highlight the unique structural properties of real-world CINs, we generated various synthetic graphs for comparison. Intuitively, using synthetic data we can generate networks that have exactly the same sizes (number of nodes), or even networks that follow identical degree distributions with our CINs. Here we use two ways to produce a synthetic network: one is to depend on an existing real network and then randomly re-arrange edges and get a randomized version of the graph, and the other is to rely on a graph generation model and produce a network with a predefined set of characteristics.

NEP: To generate a random network with the same number of nodes and edges as our initial enterprise CINs, we use the 0k-preserving re-wiring technique introduced in [8]. In a nutshell, at each iteration of the re-wiring process, one edge is selected at random and then moved to a random location, that is, it connects a different pair of nodes. This step is repeated multiple times to give a random version of the initial graph. Here, we executed 10 $|E|$ re-wirings, as suggested in [8]. In the paper, we call random networks produced by this process as Node and Edge Preserving (NEP) graphs.

DDP: For generating synthetic graphs with exactly the same degree distribution as our CINs, we used the 1k- preserving re-wiring technique from [8]. In this technique, two edges are chosen at random and then exchange node connections, that is, two given edges (v, w) and (m, n) are turned into two new ones (v, n) and (m, w). Here, we executed 10 $|E|$ re-wirings, as suggested by another work [8]. The main advantage here is that the degree distribution of the input network is exactly preserved. We use the name Degree Distribution Preserver Network (DDP) to refer to this type of synthetic networks.

BA: Here we generate synthetic graphs that follow a power-law distribution and have the same number of nodes as our CINs. Towards this end, we use the well-known Barabasi-Albert (BA) [3] generator. The generator uses the concept of preferential attachment. It keeps adding nodes, with each new nodes preferring to connect to nodes with high degree[3]. Since we want the graph to follow a power-law distribution, we do not impose any restriction on the resulting number of edges. That is, the final graph will have the same nodes as the target CIN, but potentially different number of edges.

R-MAT: Finally, we use the R-MAT generator to make synthetic graph: (a) with exactly the same number of nodes and edges as our initial CINs, and (b) that follow the structure of other popular complex networks, such as the Web graph[22]. In a nutshell, the generator uses an iterative process based on fractals [22] resulting in self-similar graphs. The model generates directed graphs, which enables a detailed comparison with our directed CINs.

IV. Identifying Communication Pattern in Enterprise CINs

Our goal in this section is to answer the following questions: (1) *How do the three enterprise CINs look and what are their differences and similarities*, (2) *How different are CINs from synthetic graphs we generated?* (3) *How is different mode communication used in the enterprise?*

A. Comparing the CALL, SMS, and EMAIL CINs

We summarize a set of scalar metrics for the three CINs in Table I, and plot six non-scalar metrics in Figure 1. As we see from Table I, all three networks are of similar scale, having 235, 234, and 151 nodes for the CALL SMS, and EMAIL CINs respectively. Even though the number of nodes is small, relatively to other complex networks, their number of edges is high leading to high edge densities ($> 12\%$). The degree distribution for all three graphs is shown in Figure 1(a). Even though there are some differences, all three CINs follow similar trends in their degree distributions.

Our next step is to compare the clustering coefficient in the graphs. In Table I, we see that CALL's average clustering coefficient is the lowest. Even though CALL and SMS are produced by the same group of employees, they have difference, indicating individuals have different communication behaviors depending on the communication channel they use. To highlight the differences, we show the distribution of clustering coefficient over all nodes in Figure 1(b). We see that more than 50% of the nodes in SMS and CALL have clustering coefficient larger than 0.5, whereas the percentage of such nodes in CALL is only 7%. This figure shows that SMS and EMAIL CINs are more similar to each other than to the CALL graph. In Figure 1(c), we show the relation between the clustering coefficient and the degree of nodes in the graphs. The x-axis shows the degree and the y-axis the mean clustering coefficient for all the nodes having that particular degree. We see that all graphs show a similar pattern, where the clustering coefficient decreases as the degree of the node increases. Moreover, we see again the similarity between the SMS and the EMAIL CINs. We plot the distribution of maximal cliques of various sizes in Figure 1 (e). In EMAIL and SMS networks, there are more maximum cliques with more nodes than in the CALL network. We speculate that we observe this because short messages and emails may be sent between a set of people with the option to reply to the entire group (e.g., reply-all in emails), but this cannot be done with phone-calls.

The degree-to-degree correlation is captured in the assortativity coefficient shown in Table I. Interestingly, the three values are around zero with values 0.095, -0.084 and -0.061, for the CALL SMS, and EMAIL, respectively. To study degree correlations in more detail, we plot the average neighbor degree metric in Figure 1(f). The x-axis shows the degree of a node and the y-axis shows the average degree of all the neighbors of all nodes with the particular degree. Unlike other complex networks, like popular OSNs, the degree of a node in CINs does not reveal information about the degrees of its direct neighbors. In other words, the average degree of the neighbors of a node u, is close to the average degree of the entire graph, and is not correlated with the degree of u. At first, this might suggest that CINs are like random networks, but it does not hold because random networks tend to have low clustering coefficient (lower than 0.15). We do not include detailed results here for brevity.

Next, we measure the network diameter and average path length over all pairs of nodes. From Table I, we observe that the diameter is 4 and average path length is close to 2, indicating that our graphs are small-world and the majority of nodes are just two hops away. The path length distribution is shown in Figure 1(d). We see that all three graphs share very similar path length characteristics.

We compare three CINs with other complex networks [2][10][4][23][7][3] and found that CINs can be distinguishable from them in terms of high clustering coefficient, edge density and around zero assortativity.

B. Comparison with synthetic networks

Here we have two goals. First, we want to compare our CINs with graphs of the same scale (e.g., same number of nodes). Second, we want to see which features of CINs are not likely to happen by chance. Such features can then be traced back to the driving forces that lead to their formation. Details on our synthetic networks are given in §III. For all our experiments, we have generated 10 random versions for each synthetic graph. Here we include results from a single run, since all versions resulted in similar results.

We have generated synthetic random networks with characteristics from all three CINs. We observed qualitatively similar observations using all three networks. For ease of exposition of the findings we only include our results from the CALL network.

We use four synthetic networks for comparison. The NEP and DPP are based on re-wiring of the CALL network, and the BA and R-MAT are produced by graph generators (see §III for details). For each synthetic network, their scalar metrics are listed in Table I and we plot their non-scalar metrics in Figure 2.

From Table I, we see that our graphs are very different from the BA power-law network. The BA network has few edges (only 372) and small clustering coefficient (0.03). The difference with BA is highlighted in the degree distribution of Figure 2(a). Moreover, the difference with BA is clearly observed in all plots of Figure 2, where the values of BA are always very far away from those of the rest of the graphs.

The R-MAT graph is closer to the targeted CINs compared to the other synthetic networks. We see the similarity in all the

TABLE I
SCALAR GRAPH METRICS FOR CALL, SMS AND EMAIL CINS COMPARED TO SYNTHETIC NETWORKS.

ID	Family	Data set	$\lvert V \rvert$	$\lvert E \rvert$	Edge Density	$\bar{k}$	r	Diameter	Path Length	$\bar{C}$
1	CINs	CALL	235	3699	13.4%	31.5	0.095	4	2	0.35
2		SMS	234	3524	12.9%	30.1	-0.084	3	1.9	0.47
3		EMAIL	151	1511	13.4%	20.2	-0.061	4	2.6	0.51
4	Synthetic	NEP	235	3699	13.4%	31.5	-0.008	3	1.88	0.13
5	Networks	DDP	235	3699	13.4%	31.5	0.009	4	1.95	0.2
6	(see §III-C)	R-MAT	234	3629	13.2%	31	0.123	3	1.93	0.21
7		BA	235	372	1.4%	3.16	-0.15	8	3.8	0.03

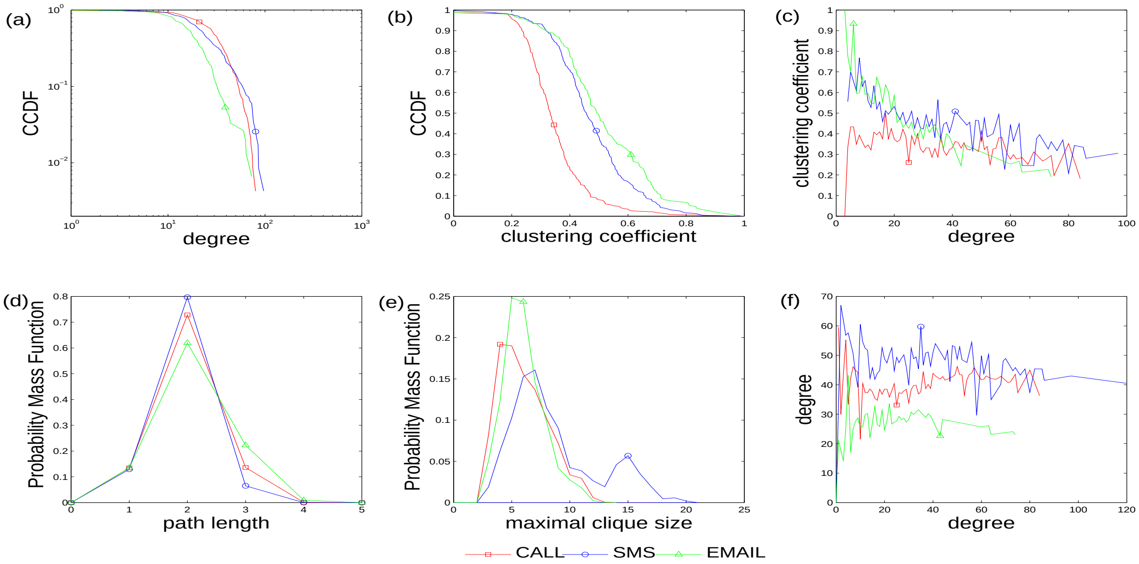
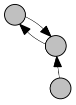
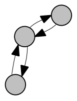
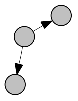
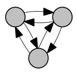


Fig. 1. Comparison of the CALL, SMS, and EMAIL CINs using six different no-scalar metrics. All networks have similarities in their degree distributions, clustering coefficient, path lengths, and their degree-to-degree correlations. The SMS and EMAIL are more similar to each than to the CALL graph.

TABLE II
PRESENTATION OF SELECTED DIRECTED MOTIFS OF SIZE THREE.

				
Motif ID	1	2	3	4
CALL	21.4%	17.2%	6.9%	3%
NEP	1.6%	0.1%	26.6%	0.004%
DPP	1.1%	0.06%	30.2%	0.001%
R-MAT	9.8%	1.5%	17.6%	0.09%

scalar metrics in Table I. From Figure 2, we see that R-MAT follows the CALL graph very closely in all metrics except the clustering coefficient. This shows that the high clustering observed in CINs is unique and hard to be captured by the R-MAT generator. We see this also in Figure 2(e) where R-MAT is not able to generate the large maximal cliques observed in the CALL network.

Our enterprise CINs are not random and they show properties that are unlikely to occur by the random connection of nodes. In particular, the degree distribution and clustering coefficient of the CALL CIN cannot be attributed to the high edge density. We show this in our comparison with the NEP randomized graph version. This synthetic network has exactly the same number of nodes and edge density as our CIN, but as we see from Figure 2(a),(b) its degree distribution and clustering coefficient, respectively, are different.

We compare the CALL graph with a DPP random graph that has exactly the same degree distribution. Even though we see from Figure 2(a) that the degree of DPP and CALL are identical, the clustering coefficient shown in Figure 2(b) is very different. Moreover, the randomized process of generating DPP fails to form the large cliques we see in the CALL network, as we observe from Figure 2(e). On the other hand, the NEP and DDP random networks have degree-to-degree correlations (e.g., assortativity) and path lengths (Figure 2(d)) that are very similar to the CALL graph's.

Our findings indicate that the clustering of the CIN graphs is the main property that distinguishes them from the synthetic graphs. To further highlight these differences, we decompose these networks to their corresponding sub-graphs (a.k.a. motifs [13]) of size three. We summarize motifs we found particularly interesting in Table II. Motifs 1,2, and 4 occur with much higher frequency in the CALL graph, compared to the synthetic graphs. In particular, the directed full-way clique-motif (4) appears two orders of magnitude more often

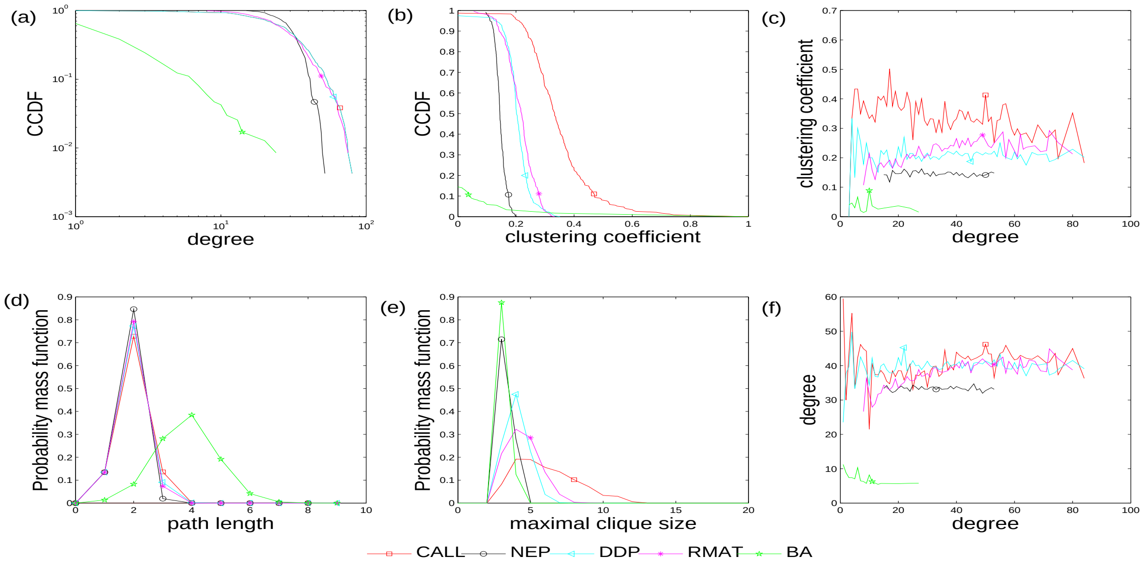


Fig. 2. Comparing the CALL network with four synthetic graphs. The CALL CIN is different from random graphs, especially in the clustering and the formation of large cliques. From all the synthetic graphs, the power-law (BA) is the most different from our CALL CIN.

than in any of the synthetic networks. This reflects the nature of CALL networks that edges represent human interactions, so interactions are usually bi-directional. For example, the motif 3 shows two edges that are not bidirectional. This motif happens very frequently by chance, but in CINs its frequency is considerably lower. The findings hold in all three CINs.

C. *Communication patterns in CALL and SMS*

A unique advantage of our dataset from the communication provider is that for the same group of employees we have two different CINs, one using short message and one using phone calls. This enables us to study the behavior of individual employees when using these two different communication modes. In this section we answer the following questions: (a) *Are the local graph features (e.g., the degree) of an employee in CALL similar to his or her corresponding characteristics in SMS?* (b) *Are the employees likely to use single-mode or multi-mode communications?*, and (c) *Are there any differences between ordinary employees and managers?*

To answer the first question, we use the scatter-plots in Figures 3 (a) and (b) where we compare the degree and clustering coefficient for each employee in the two graphs. In both scatter-plots we see that some points are far away from the diagonal, showing differences in how the two communication modes are used by some employees. On the other hand, as a general trend, both plots indicate a loosely positive correlation, where if a node has high node degree or clustering coefficient in one network, it will have high degree of clustering coefficient in the other network. This supports our observation in the previous section where the two CINs showed similar structural graph properties.

An interesting observation from Figure 3 (a) is that 40% of the points are above and 60% below the diagonal. This shows that most employees have a larger local neighborhood size in the CALL compared to the SMS graph. On the other hand, the clustering coefficient scatter-plot (Figure 3 (b)) shows a different trend, where 85% of the points are now above the diagonal. This shows that even though most employees have smaller degree in the SMS CIN their local neighborhood is on average better connected. This observation is suggestive of a collaborative behavior where employees tend to use short messages to communicate with their close collaborators forming well connected neighborhoods. In the next section, we show results where the higher clustering coefficient of the SMS graph can explain why the SMS CIN is better for inferring hierarchies compared to CALL.

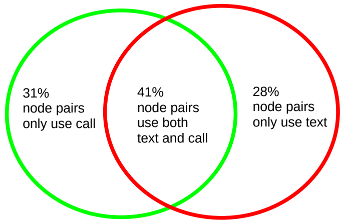


Fig. 4. A Venn diagram comparing the number of communicating pair of nodes that use: (a) only short messages, (b) only phone calls, and (c) both short messages and phone calls. From the diagram we see than most of the communicating node pairs use a single-mode communication and only 41% are observed to use multi-mode communications.

For the second question, we find 2056 communication node pairs to be shared by both CALL and SMS, while 1634 and 1468 communication node pairs exclusively belong to CALL or SMS respectively. Our findings are graphically illustrated in Figure 4. From the figure we see that the corresponding percentages are 41%, 31% and 28%, which are very close to each other. In other words, if we randomly select a pair of communicating employees, is equally likely to use only short message, only phone calls, or both. Interestingly, it is more

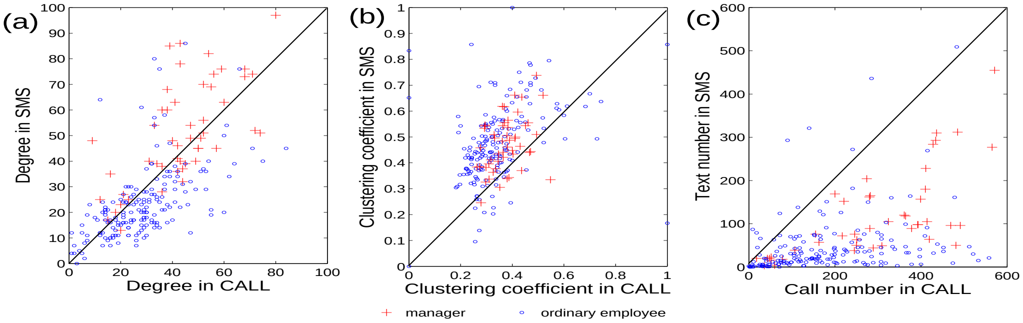


Fig. 3. In the scatter plots we compare the characteristics of each employee in the communication provider's enterprise using the CALL (x-axis) and SMS (y-axis) CINs. In (a) and (b) we compare the degrees and clustering coefficients for the managers (crosses) and ordinary employees (dots), for the CALL and SMS networks respectively. In (c) we compare the number of phone calls made and short messages send by each employee over the duration of our dataset.

popular in this enterprise for employees to use a singe-mode than a multi-mode communication. This observation indicates that when studying interaction within an enterprise adding more communication modes increases the number of observed links in the interaction graphs.

For the third question, we plot total number of calls made by an employee against his or her total number of short messages. We illustrate these results in the scatter-plot of Figure 3 (c). We see that most points are below the diagonal, showing that employees make more phone calls than text messages. It is very interesting to observe that managers are closer to the diagonal line than ordinary employees. This indicates that managers use both phone call and text to contact others in a relatively balanced way. In contrast, ordinary employees are inclined to use phone call as their priority choice. These observations motivate our next section where we utilize the information within CINs in order to infer the roles of different employees.

V. Detecting Hierarchy Structure in Enterprise CINs

In this section, our goal is to use CINs to infer the position of employees in the hierarchy of the enterprise. Moreover, we want to see if the choice of CIN (e.g., CALL versus SMS) can affect the inference process.

A. *HumanRank: Ranking employees in enterprises using CINs*

Any ranking method works as follows. First, it assigns a score to each node (employee) in the graph (enterprise). Then, it ranks all nodes by sorting them according to their assigned scores. The node with the highest score gets rank=1 and the one with the lowest gets $|V|$. In our case, our goal is to give a score to each node that describes its status in the enterprise.

Our ranking method is motivated by the following observation: The importance of an individual can be inferred from the importance of the people he or she interacts with. Intuitively, senior managers will interact more other managers than ordinary managers. A similar observation regarding the structure of the Web lead to Google's PageRank [14]. In PageRank, if a web page is linked from other important web pages, it is regarded as a more important page.

In this section, we introduce **HumanRank**: An iterative process of ranking individuals based on their position in the hierarchy. We have adapted PageRank to fit the requirements of our problem. The PageRank algorithm treats edges differently based on their direction and one of its simplied versions for calculating v's PageRank is defined below:

$$PR(v) = \sum_{w \in B(v)} PR(w)/D(w)$$

where $B(v)$ contains all pages linking to page v and $D(w)$ is the outdegree of w.

In our work, we treat the graph as undirected. In the Web, if a small website w has a hyperlink to a popular website (e.g., google.com), this does not say anything about the importance of website w. However, if an individual contacts the CEO of a company, there is a high chance that this individual is also important. We also applied PageRank without our adaptation and observed lower accuracy in all our directed graphs. In HumanRank the hierarchy score of a node is defined as follows:

$$H(v) = \sum_{w \in L(v)} H(w)$$

Where $L(v)$ is the set of v's neighbor nodes. The complete process is outlined in Algorithm 1. In pratice, the iteration process stops when the scores of each node does not change significantly from the previous step. In all our experiments, we used a 1% threshold, which resulted in convergence after 10 to 20 iterations

Exploring other ranking methods. We are interested in the following question: *Why not use a simple graph metric for ranking, such as the degree or the clustering coefficient of a node?* As we see from Figure 3 (a), the degree of a manager is not absolutely higher than an oridnary employee, especially in the CALL network where the average degree of ordinary managers (43.9) is higher than that of senior managers (38.4). The clustering coefficient also shows its limitations. These observations highlight the difficulty of the problem at hand and suggest that trivial solutions are unlikely to lead to good results. We revisit the comparison with other methods in §V-C.

B. *Hierarchy detection using HumanRank*

Given the ranking results from HumanRank, we want to form two groups. Intuitively, nodes with low HumanRank scores are more likely to be ordinary employees, whereas those with high scores are more likely to be managers. In

Algorithm 1 HumanRank

```
Require: a network $G$
  for each $v \in V(G)$ do
    $H(v) \leftarrow 1$
  end for
  while $H(v)$ changes do
    for each $v \in V(G)$ do
      $NH(v) \leftarrow \sum_{w \in L(v)} H(w)$, where $w$ is one of $v$'s
      neighbor nodes
    end for
    $\overline{NH} \leftarrow \sum_{v \in V(G)} NH(v)/|V(G)|$
    for each $v \in V(G)$ do
      $H(v) \leftarrow NH(v)/\overline{NH}$
    end for
  end while
  return $H(v), v \in V(G)$
```

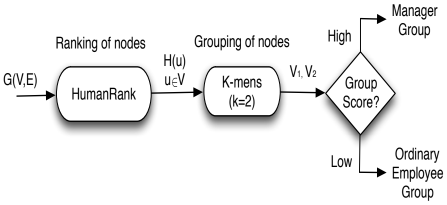


Fig. 5. The two step process for labeling nodes as ordinary employees or managers. All the nodes are firstly ranked according to HumanRank and then are assigned in two groups using k-means (k=2).

addition, we want our classifier to be unsupervised. That is, we do not want to depend on training data, which are often hard to obtain in practice. We compare our approach with supervised machine learning classifiers in §V-C.

The steps of our classification algorithm are outlined in Figure 5. First, we calculate HumanRank on the input CIN. Next, we use K-means to group nodes into two disjoint groups based on their HumanRank score. The last step is to label each cluster as ordinary employees or managers. We use the average ranking score of each group to achieve this. All the nodes in the cluster with the high score are labeled as managers. Similarly, all the nodes in the other cluster are labeled as ordinary employees.

C. Experimental Evaluation

In our dataset, we have the job description of every employee, therefore we use this information as **ground truth** to evaluate our algorithms. The classification **accuracy** is defined as the number of correctly classified employees divided by the total number of employees with ground truth. **F-score** is a measure of a test's accuracy and considers both precision and recall.

1) Ranking results: For evaluating the ranking results, and comparing it with other methods, we use the following test. We rank nodes using their score by a given method, e.g., HumanRank. We then measure how many managers are located in the top k ranked nodes and measure this for different k's in the range $1...|V|$. We compare our results with the state-of-the-art in hierarchy detection, which uses graph entropy [1]. In addition, we compared with ranking based on the node degree and PageRank. We also used ranking on clustering coefficient that did not give good results and we omit it from the remaining of the paper.

The comparison results are illustrated in Figure 6. We also report the results of an Optimal ranking method, which always gives better ranking to managers than ordinary employees. As we see from the figure, the HumanRank gives results that are closer to the optimal ranker than others. In particular, we are doing significantly better than the state-of-the-art method, which uses entropy [1]. In Figure 6, we magnify the ranges up to rank 20 for better comparison. In the top 10 ranks for all three CINs, our approach gives eight, nine, and ten managers, which corresponds to 90% of the time on average. This is in contrast to 70%, 65%, and 67% for the degree, PageRank, and entropy rankings, respectively.

It is also interesting to observe that the ranking results are better when we use the SMS network compared to the CALL network, even though they capture interactions from the same enterprise. This suggest that the “informal” exchange of SMS can reveal the hierarchical structure of an enterprise better than the more “formal” exchange of phone-call. In the EMAIL network, the ranking appears more challenging for all methods. However, even with this data set, our algorithm gives better results over a long range of ks. We hypothesize that the lower accuracy in the Enron dataset is due to the removal of some email exchanges after personal requests by the participating individuals [24]. Since retrieving the lost emails is hard, establishing causality can be difficult.

2) Hierarchy detection: We group the competing methods into supervised and unsupervised solutions. Supervised solutions combine graph features for different nodes. We tried different machine learning methods: logistic regression, SVM, Random forest, and Bayesian Networks. The features we used for classification are: degree, clustering coefficient, betweenness, average degree of its neighbors, and the average distance to all other nodes in the graph. We then use 10-fold cross validation to generate the results in the paper. That is, we used 90% of the data for training and 10% for testing. We repeat this several times until all the instances are included in the testing and training at least once. For the unsupervised methods, we repeat the same process as in our classifier, but we rank nodes using the method in [1].

We applied our classification results on all three CINs. In Figure 7 (a) and (b), we show the accuracy of our approach compared to different methods. As we see, our method gives consistently good results. Even though in some data sets another method can be better, we found the performance of these approaches to vary significantly across different networks. Especially, HumanRank outperforms other methods in each CIN in terms of manager identification, which is an important advantage of HumanRank, because inferring key members in an organization is more meaningful than inferring ordinary members.

VI. Summary and Conclusions

With our work we aim to give an in-depth study of the communication patterns formed by employees interacting in two modern enterprise organizations. We call these graphs Communication Interaction Networks, or CINs. Towards this end, we use data from two different organizations; the Enron email dataset, and data from the exchange of short messages and phone-call from a communication provider. Our data are unique in two ways: (a) we are the first to study CINs

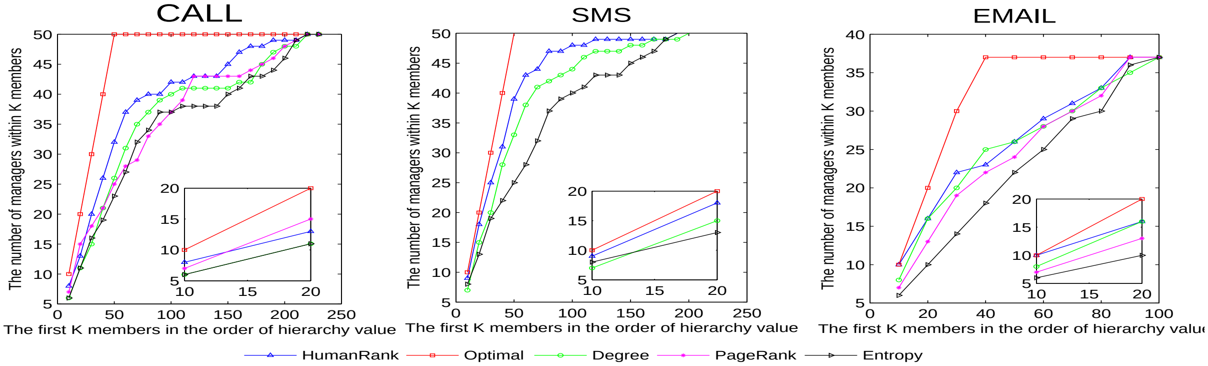


Fig. 6. Ranking results: The plots show the number of managers identified in the top k nodes, as ranked by our HumanRank and three other ranking methods. We also compare our results with an optimal ranker where all managers have higher score than ordinary employees. PageRank is not applied on SMS because SMS is undirected.

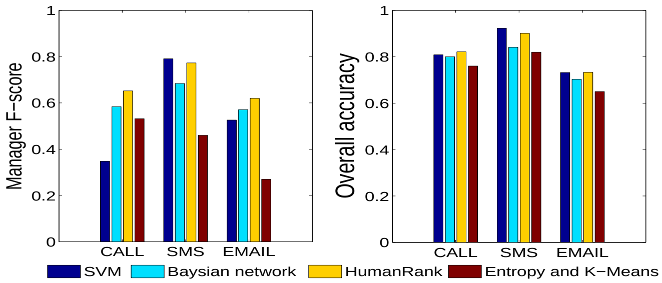


Fig. 7. Accuracy and Manager's F-score in hierarchy detection using HumanRank and K-means (k=2). We also show measure results of two supervised learning methods, and one unsupervised method using the ranking from [1] and K-means (k=2).

using two different organizations, and (b) we are the first to study two CINs by the same employees and two different communication modes. The key observations from our study are: (a) three CINs share common features despite of their diverse background; (b) employees that play different roles have distinct communication behavior. (b) CINs carry valuable information about the hierarchical structure of an enterprise, and our HumanRank method can correctly identify the job titles for above 70% of their employees.

We hope that our observations and proposed algorithms can find their way to other applications that use interaction networks. For example, it will be very interesting to see if we can identify leaders in criminal or terrorist organizations using only the observed interactions.

References

[1] J. Shetty and J. Adibi, "Discovering important nodes through graph entropy the case of enron email database," in *SIGKDD*, 2005, pp. 74–81.

[2] M. E. J. Newman, "Scientific collaboration networks. i. network construction and fundamental results," *Phys. Rev. E*, vol. 64, p. 016131, 2001.

[3] A. L. Barabasi and R. Albert., "Emergence of scaling in random network," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[4] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *WWW*, 2007, pp. 835–844.

[5] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in facebook: A case study of unbiased sampling of osns," in *INFOCOM*, 2010.

[6] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *SIGCOMM*, 1999.

[7] H. Jeong, S. Mason, A. L. Barab'asi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, pp. 41–42, 2001.

[8] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat, "Systematic topology analysis and generation using degree correlations," in *SIGCOMM*, 2006, pp. 135–146.

[9] D. Reinhard, *Graph Theory*, third edition ed.

[10] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, pp. 400–442, 1998.

[11] M. E. J. Newman, "Assortative mixing in networks," *Phys. Rev. Lett*, vol. 89, p. 208701, 2002.

[12] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, pp. 163–177, 2001.

[13] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[14] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web," *Technical Report, Standford InfoLab*, 1999.

[15] A. Agarwal, S. Chakrabarti, and S. Aggarwa, "Learning to rank networked entities," in *SIGKDD*, 2006.

[16] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[17] D. Chen, J. Tang, J. Li, and L. Zhou, "Discovering the staring people from social networks," in *WWW*, 2009.

[18] K. Thomas and V. Milan, "Behavioral profiles for advanced email features," in *WWW*, 2009, pp. 711–720.

[19] R. Rowe, C. Creamer, S. Hershkop, and S. J. Stolfo, "Automated social hierarchy detection through email network analysis," in *In Joint 9th WEBKDD and 1st SNAKDD workshop*, 2007, pp. 109–117.

[20] J. Leskov, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *SIGKDD*, 2005.

[21] J. Abello, P. Pardalos, and M. G. C. Resende, "On maximum clique problems in very large graphs," *DIMACS Series on Discrete Mathematics and Theoretical Computer Science*, vol. 50, pp. 119–130, 1999.

[22] D. Chakrabarti, Y. Zhan, and C. Faloutsos, "R-mat: a recursive model for graph mining," in *SDM*, 2004.

[23] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. J. Shenker, and W. Willinger, "The origin of power laws in internet topologies revisited," in *In Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies*, 2002.

[24] G. Carenini, R. Ng, and X. Zhou, "Scalable Discovery of Hidden Emails from Large Folders," in *SIGKDD*, 2005.