Enterprise Social Network Analysis and Modeling: A Tale of Two Graphs

Jin Cao[1], Hongyu Gao[2], Li Erran Li[1], and Brian Friedman[1]
[1]Bell Laboratories, Alcatel-Lucent, USA; [2] Northwestern University, Evanston, IL, USA

Abstract—**Like their public counterpart such as Facebook and Twitter, enterprise social networks are poised to revolutionize how people interact in the workplace. There is a pressing need to understand how people are using these social networks. Unlike the public social networks like Facebook or Twitter which are normally characterized using the social graph or the interaction graph, enterprise social networks are also governed by an organization graph. Based on a six month dataset collected from May through October 2011 of a large enterprise social network, we study the characteristics of activities of its enterprise social network. We observe that the user attributes in the organization graph such as geographic location (eg. country) and his/her rank in the company hierarchy have a significant impact on how the user uses the social network and how user interacts with each other. We then build formal statistical models of user interaction graphs in enterprise social network and quantify effects of user attributes from organization graphs on these interactions. Furthermore, as the enterprise social network medium bring users from diverse locations and social status forming ad-hoc communities, our statistical model can be further enhanced by including these ad-hoc communities.**

I. Introduction

Public online social networks (OSNs) have gained tremendous popularity and have already revolutionized the way many people communicate. With the success of public OSNs like Facebook and Twitter, employees, especially young employees, have experience with OSNs as tools for communication. Companies are quick to use public OSNs to interact with customers. However, most of them do not want to use public OSNs for internal collaboration and communications. A recent survey [10] shows that 71% of companies block public OSNs in their workplace. The main reason is that such OSNs typically sit outside the corporate firewall, and that users may inadvertently reveal sensitive information to outsiders.

On the other hand, traditional forms of communication using phone, email, and instant messages are woefully inadequate in today's enterprises, which must respond to customer needs quickly and increasingly coordinate among geographically diverse sites. To address the needs of companies, enterprise social network software (or internal social network software) have been developed, e.g. Salesforce.com's Chatter and Jive Software [8]. Using such software, enterprises can deploy internal social networks that are used within the enterprise boundary, protected behind company firewalls. The access to the content in enterprise social networks is restricted to employees, while within the enterprise boundary itself, a significant proportion of the content is public. These enterprise social networks have the potential to drive knowledge worker efficiency to new levels; they are incredibly effective at allowing the efficient exchange of knowledge and expertise across geographic and organizational boundaries that have traditionally stifled knowledge capture and sharing [14].

One important distinction between enterprise social networks and public OSNs is that users of the former (employees) are not equivalent peers. Rather, the corporation has a tree-like hierarchy, where each individual user resides at a certain position (or level). We refer to this structure as the organization graph. The resulting organization graph imposes certain relations between the interacting user pairs in the social graph (an edge exists if the two users communicate), for example, manager-subordinate relation, coworker relation, *etc*. It is unclear how the corporation hierarchy affects user interaction in the social network. Another important difference is that every employee can potentially interact with every other employee. In contrast, users in public OSNs do not interact with other users not in their acquaintance list.

Enterprise social networks are adopted by between 8% and 40% of companies [5]. Their adoption continues to rise and is forecasted to continue rising for years. There is a great need to understand how enterprise social networks are changing employee interaction within enterprises. For example, to what extent do enterprise social networks break organizational and geographic boundaries, and bring employees together as a tightly-knit community? Do enterprise social networks improve the effectiveness of communication between company management team and employees? We are not aware of any quantitative studies. In this study, we analyzed an enterprise social network and its correlation with the corporation hierarchy structure for a large international corporation with about 79 thousand employees. The social networking system is based on Jive Software [8], which has received the highest overall score from Forrester [14] among all such applications. In this paper, we will simply refer the enterprise social network used within the corporation as *Jive*. Although the Jive system was officially launched within the corporation about 2 years ago, it has become increasingly popular recently since it has been promoted as the preferred form of interaction and its adoption is considered a success [6].

Our study was based on 6 months' worth of crawled Jive data obtained from May through October 2011. Our data contains about 56,000 activities generated by 7,400 unique users which is about 10% of company employees. Note that since the Activity Crawler is only able to obtain data from users that post public and members-only data to Jive, we have no way of knowing how many users may be posting only to private or secret groups, nor how many users view Jive content but never post anything at all.

We make the following contribution in this paper.

- To the best of our knowledge, we are the first to study the influence of different relations (*e.g.* , coworker, supervisor, subordinate) between nodes on their interaction in an online enterprise social network in large scale. Our data provides a *complete* view of interaction within the network in the data collection period, instead of a sampled subset of it. Thus, it serves as a reliable foundation to draw all our further findings on.
- We propose a formal logistic regression model to quantify the effect of user's relations on their interaction patterns in the enterprise social network. Although in this study we only examine the effect of user relations extracted from the organization graph, including location, coworker relation and supervisor-subordinate relation, our methodology can easily apply to other relations.
- We make a series of interesting observations through the analysis. For example, both user's geo-location and position in corporate hierarchy are highly significant in predicting their interactions. As another example, the enterprise social network medium brings together users from diverse locations and social status forming ad-hoc communities. Including these communities in the statistic models improves the fit significantly. Influential users are distributed across different ranks (tend to be higher) and in different communities. The two observations do not contradict each other. This is because the number of users adjacent to a particular user in the organization graph is small as comparing to users that are far apart. So for users that are far part, even though the probability of their interactions is small, we still observe many occurrences of such interactions.

The majority of this paper focuses on analyzing and modeling user interaction graphs in enterprise social networks. For privacy concerns, we do not present any analysis on the actual contents of the enterprise social activities.

The remainder of this paper is organized as follows. In Section II, we provide necessary background information about the Jive social network as well as our data collection approach. Next, in Section III, we describe the enterprise social interaction graphs in Section III and present an analysis of the corporate hierarchy attributes in Section IV. We then model the impact of corporate hierarchy on the interaction graph in detail in Section V. Lastly, we review related work in Section VI and conclude the paper in Section VII.

II. Data

A. Enterprise Social Network

The enterprise social network application that we employed in our study is a web-based system using the Jive software. It has been used for approximately two years by a large international corporation under study with about 79,000 employees. All activities in the social network are by default open to any user, unless specifically set to be private. Similar to Twitter, users can follow other users to get a convenient feed of the updates of their followings. The Jive system currently supports four basic types of activities, each designed for a different situation: "document", "blogpost", "discussion" and "microblog." A "document" is a more formal piece of content that typically consists of an uploaded file along with an associated description. Several authors may have worked together on the document, but only one typically uploads it to Jive. Other Jive users can submit comments about the document and, unlike the other activity types, documents can be rated by users on a scale of 1 to 5. A "blogpost" is created by a single author to share one or more ideas, typically with lots of detail. Although other Jive users can reply to a blogpost entry, the purpose of blogposts is more informational and less about interactions with other users. A "discussion" is more suitable for sharing an idea or asking a question where replies are welcomed and/or expected. A "microblog" is equivalent to a status update in Facebook or a tweet in Twitter, and is typically short. Other users can post comments about the microblog entry.

The Jive instance within the enterprise we study provides four data privacy levels: public, members-only, private and secret. Due to data privacy constraints, we can only access the public and members-only data, which we are told represents approximately 50% of the data being posted. In the Spring of 2011 we began development of an "Activity Crawler" which used a few of Jive APIs to extract activity data from the system. By May 2011, our Activity Crawler was executing autonomously, once per hour, to retrieve public and members-only activity information from Jive, and storing it in a MySQL database. We chose the 1 hour frequency to balance the need for up-to-date Jive activity information without placing too much load on Jive itself. For each collected activity we record the $objID$, $parentID$, $empID$, $type$, and $timestamp$. The $objID$ (in combination with the $type$) uniquely identifies the activity. The $parentID$ is either the $objID$ of the activity that the current activity replies to, or is identical to the $objID$ if the current activity is new. The $empID$ identifies the employee that generated the activity. As mentioned above, there are four $types$: document, blogpost, discussion and microblog. Each activity is further classified as being a new posting, a reply to a previous posting, a modification to a previous posting, a move of a posting (and its related replies) from one "group" to another, or a user rating of a posting.

From May through October 2011, the crawled data contains about 56,000 activities generated by 7,400 unique users. These does not include postings only to private or secret groups and passive user activities such as content viewing.

B. Corporate Organization Graph

Distinct from the social networking application, the company also provides an internal web interface to query any individual employee. The result reveals the employee's location, manager as well as all of that manager's subordinates. Starting with 5 random employees, we used this interface to traverse the tree structure of the company's organizational hierarchy. We collected information for about 62,000 employees in this way. We could not collect information for all employees mainly because the query system is not perfect. In addition, we developed an algorithm to determine the country in which each employee resides based on the available employee data, including address and work phone number. We were able to successfully determine the country of origin for more than 99% of the 62,000 employees. We also recorded the $empID$ and his/her manager's $empID$ for all the employees collected.

III. Interactions in Enterprise Social Network

As described in Section II-A, Jive users can either create a new piece of content or act on an existing piece of content. When a user responds to an existing piece of content, either through a modification, comment, reply or rating, an instance of *interaction* is created. In this section, we present our first analysis of the interaction graphs between Jive users in the enterprise social network. We focus primarily on usage patterns and basic graph properties.

A. Activities Breakdown

There are four basic types of content a user can contribute in the Jive system (Section II-A). Among the collected activities, "document", "blogpost", "discussion" and "microblog" take up 34.1%, 19.5%, 40.0% and 6.4% in number of activities, respectively. It is apparent that microblog is less used than the other three types. One possible reason is that the company also has provided a completely separate enterprise-wide microblogging solution nearly two years prior to the launch of the Jive system. Many company employees continue to use the legacy microblogging system rather than the microblog functionality within Jive.

We further break down the usage of each activity type into five subcategories: creation, modification, comment/reply, move and rating. We observe that documents and blogposts are modified heavily after their creation. This is due, at least in part, to users saving their work more than one time during the creation of the document or blogpost (each subsequent "save" is registered by the application as a "modification" to the document or blogpost). The number of "modification" activity reaches 27.3% of the total activity numbers in these two categories. It is intuitive since they are designed for more formal interaction with lots of detail. In contrast, this ratio is only 0.3% and 1.7% for discussions and microblogs, respectively. In addition, one discussion on average receives 3 replies. It reflects that the "discussion" type does successfully trigger discussions among the employees.

B. Interaction Graph Analysis

Type	#Nodes	#Directed Pairs	#Undirected Pairs	#Act
All	8808	30490	24563	47097
Discussion	6938	23243	17700	33834 (72%)
Document	3383	4846	4736	7534
Blogpost	2190	3088	3037	4441
Microblog	257	614	543	1288

TABLE I
Sizes of Interaction Graphs

1) Interaction Graph Generation: We construct graphs based on Jive users' interactions on the enterprise social network, whereas the Jive users are the nodes, and each interaction creates a directed edge from the author of the new activity to the author of the original piece of content. We refer to such a graph as *interaction graphs*. In the graph generation process, we removed self-loops where users interacted with themselves, removed all isolated users with no interactions, and merged multiple edges with the same interacting user pair into one edge with a corresponding weight. We generated an overall interaction graph containing all four activity types, as well as one graph for each activity type, thus creating five interaction graphs in total. Table I shows some general statistics about the five graphs. Notice that the total activity number is only 36% of the total activities reported in Section II-A, since the graph does not contain any self-loops or activities that do not trigger interactions.

An interesting discovery from Table I is that the number of edges in the overall interaction graph is approximately the same as the sum of the edge numbers of the four activity-specific interaction graphs, reflected in both column 2 and column 3. This means that the four activity-specific interacting graphs have mostly disjoint edges. The implication is that for the majority of interacting user pairs, they only interact in the context of one specific type of social activity. For example, a pair only uses "discussion" to interact, but never uses "document", "blogpost" or "microblog".

2) Graph Statistics Summary: Table II shows some basic properties of the five interaction graphs, and we observe a few interesting facts: i) "discussion" and "microblog" interactions have a much higher probability of being reciprocal than "document" and "blogpost" interactions, probably due to the fact that "discussion" and "microblog" are more *casual* interactions; ii) There are giant connected components in all graphs which include more than 80% of nodes; iii) The clustering coefficients are much lower than the reported number of public social networks (0.167 for Facebook [13] and 0.330 for LiveJournal [11]), except for "microblog" which is similar to activities in public social networks like Twitter. This may imply that the nature of interactions in "discussion", "document" and "blogspot" is different from public social networks. For example, before creating a "document" in the enterprise social networks, the users may have already exchanged ideas in a meeting, but interactions which occurred during the meeting would not be captured in the social network usage; and iv) All graphs have small diameters (less than 20).

Type	Reciprocity	Giant Component	Cluster Coefficient	Diameter
All	19%	93%	0.06	14
Discussion	28%	89%	0.07	16
Document	2%	80%	0.04	18
Blogpost	1%	84%	0.02	14
Microblog	16%	90%	0.17	7

TABLE II
Graph Statistics of Interaction Graphs

IV. Organization Graph As Attributes of Interaction Graphs

Jive users reside at certain positions in the corporate hierarchy. For a pair of users, their distance in the corporate organization graph may impact their behaviour in the social graph, for example, a pair with manager-subordinate or coworker relation may interact more in the social graph comparing to an arbitrary user pair. In this section, we analyze *empirically* how a user's attributes (geographic location and position) in the company hierarchy impact his/her behaviour in the Jive social graph.

For a Jive user, define his/her *hierarchy level* as the number of hops to the top level of the corporate hierarchy. For a Jive user pair, define their *hierarchy distance* as the number of hops to their nearest common ancestor in the corporate hierarchy, whichever is larger. Our main observations are as follows: i) Jive users with interactions tend to have a lower hierarchy level and have a smaller hierarchical distance relative to the whole employee population, ii) Enterprise social networking tools such as Jive truly stimulate the interactions between employees that are further apart in the corporate hierarchy (further than direct boss and peering relationship), iii) How well a user is responded to in the enterprise social networking activities is positively correlated with how influential he is in the company. The first observation is even more true for important users in the enterprise social networking activity (measured for example by the betweenness metrics).

A. Distributions of Organization Graph Attributes

Figure 1 shows the histogram of hierarchy levels of users in the interaction graph. As a reference, we also plotted the corresponding histogram for all employees (dashed). It is obvious that, in comparison, the Jive users in the interaction graph tend to sit higher in the corporate hierarchy.

Figure 2 shows the histogram of hierarchy distances for all interacting user pairs (solid). The hierarchy distances for interacting user pairs heavily distribute in the region from 3 to 7. Recall that a distance of k means that the user pair have to go up in the corporate hierarchy for k hops to reach the nearest common ancestor. It indicates that enterprise social networking tools such as Jive truly stimulate the interactions between employees that are far apart in the corporate hierarchy. We also compare the distribution with the corresponding distributions for an arbitrary Jive user pair (dotted) and an arbitrary employee pair (dashed), by randomly selecting 1000 Jive users and employees. It is obvious that the interacting pairs typically tend to have smaller distances in the hierarchy compared to the other two. This is expected since the interactions in enterprise social network are mostly work related.

B. Organization Graph Attributes for Significant Nodes and Communities

We study the influential users (nodes) and communities in the interaction graph in terms of their corporate hierarchy attributes. For this, we use the interaction graph for all types of interactions and treat it as an undirected graph.

We use betweenness centrality as a measure of user significance. Let v be an arbitrary user (node) in the interaction graph, then its betweenness centrality is given by the expression:

$$g(v) = \Sigma_{s \neq v, t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}}, \qquad (1)$$

where σ_{st} is the total number of shortest paths from node s to t and $\sigma_{st}(v)$ is the number of those paths that pass through v. Note that the betweenness centrality of a node scales with the number of pairs of nodes as implied by the summation indices. Using betweenness centrality (1) as a measure of significance, we discovered a total of 27 users that have a betweenness value much higher than the others. We, thus, consider them as the *influential users (nodes)*. There is at least one influential user for 9 out of the top 10 countries with the largest number of users, with the top 2 countries having around 56% of influential users. Interestingly, these users also tend to sit higher in the corporate hierarchy (see Figure 1) comparing to the Jive users and all employees. The clustering coefficient for the subgraph formed by these 27 users is much higher with a value of 0.61, indicating that they form a closely intertwined community even though they are geographically dispersed.

We also study the distribution of organization graph attributes for communities formed in the interaction graph. The communities are discovered using the leading eigenvectors of the graph modularity matrix [4]. Figure 3 shows the discovered 6 communities in the interaction graph formed by 542 users (out of 5742) with more than 10 interactions. On the figure, we also marked by stars where the influential users are w.r.t. these communities. It is obvious that these influential users are distributed across communities. A close look also reveals that each community is composed of users that are geographically dispersed. Later in Section V, we use communities like these to aid the modeling of interaction graphs.

C. Effect on User Interactions

We are interested in studying how attributes derived from organization graph affects the user interaction behaviour.

We first study how a user's hierarchy level affects the number of his/her interactions, i.e. the degrees. For a particular user, let the in-degree be the number of other users that have replied to him, and the out-degree be the number of other users that he/she has replied to. The correlation between the in-degree and the hierarchy level has an estimated Pearson correlation coefficient of -0.23 and is statistically significant (p-value < 0.001). (It should be noted that p-values are computed based on the asymptotic Gaussian distribution assumption when the number of observations are large which is the case here.) Our result implies that the users that reside higher in the company hierarchy tree are more likely to receive replies from other users, which is intuitive since they are more influential people in the company and receive more attention. However, for out-degree, the Pearson correlation coefficient is -0.01 with a p-value of 0.95 (not statistically significant). This indicate that the hierarchy level does not affect the user's outgoing behavior significantly.

To study the effect of a user's geographic location on degree, we computed an average degree for each user in the top ten countries. Our computations indicate that developed countries (with a mean degree 2.8) have higher degrees than developing countries (with a mean of 1.7).

V. Modeling Social Interaction Graphs Using the Organization Graph

In this section, we *quantify* the effect of a user's position in the organization graph on his/her social interactions in the Jive

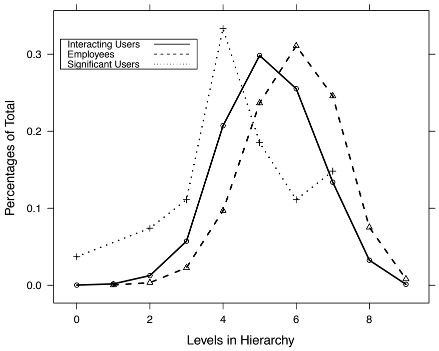

Fig. 1. Histogram of the hierarchical level of users in interaction graph (solid) and of employees (dashed), and 27 influential users that have a betweenness centrality value much higher than the others in the interaction graph (dotted) (details described in Section IV-B) .

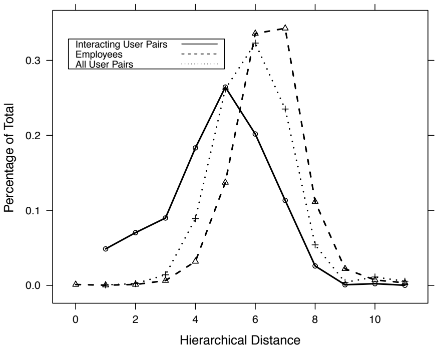

Fig. 2. Histograms of hierarchical distances for Interacting user pairs (solid), any employee pair (dashed), and any Jive user pair (dotted).

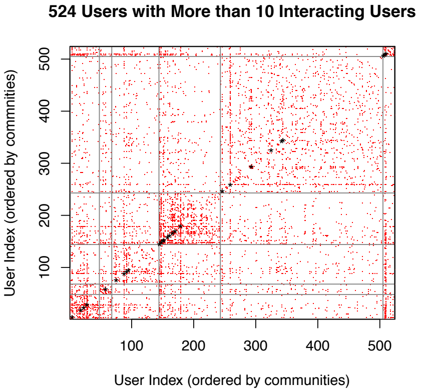

Fig. 3. The adjacency matrix of the user interaction graph, with users ordered by the six communities. The lines show the boundaries of communities. Stars in diagonal marks the 27 influential users in the graph.

enterprise social network through *formal statistical modeling*. This provides us a deeper understanding of the interactions between the organization and Jive social graph.

To simplify the modeling task, we treat the user interaction graph in the enterprise social network as an undirected graph. More specifically, a pair of nodes are considered as connected by an undirected edge if there is a directed edge between them in either direction. Moreover, we do not consider the weight of the edges which represents the number of interactions between the same user pair.

While strictly speaking, both of the graphs (interaction and organization) evolve over time, the organization graph evolves at a much slower pace (no major events occurred during the 6 months of data in this study). For further simplification, we treat the organization graph as a static graph and use it to model the interaction graph built from the entire 6 months' worth of data.

In the following, we first layout the logistic regression modeling framework for interaction graphs, and discuss general issues on model fitting and significance test. Then we present a few sets of model used in our study by incorporating different aspects of organization graph attributes as covariates. Next we present analysis results and discuss their implications. Finally, we propose an enhancement of the model by adding the latent communities as additional attributes.

A. Logistic Regression Modeling Framework

We model the user interaction graph as a random graph, meaning that it is generated by a random process. We model edges in the user interaction graph as Bernoulli random variables, which takes a value of 0 or 1 with a certain probability value. Let N be the total number of users. For a pair of users $(i, j), i, j = 1, \ldots, N, i \neq j$, let Y_{ij} be the indicator variable of the presence of an interaction between the user pair. Then Y_{ij} is modeled as a Bernoulli random variable with probability p_{ij}, i.e.,

$$\mathcal{P}(Y_{ij}) = p_{ij}^{Y_{ij}}(1-p_{ij})^{(1-Y_{ij})}. \tag{2}$$

We further make a simplifying assumption by treating $Y_{ij}, i, j = 1, \ldots, N$ as independent random variables. Later on (Section V-A2), we will comment on how to relax this assumption and allow dependencies between interactions using the same modeling framework. Notice that when p_{ij} are all equal (independent of i, j), this is the well-known Erdös-Rényi model in graph theory. In contrast to constant p_{ij}, the main focus of our study is to model how p_{ij}, the propensity of a connection between a user pair, is affected by their mutual relationship in corporate hierarchy.

For the user pair (i, j), let X_{ij} be a set of covariates derived from their relationship in the organization graph. Furthermore, let Z_{ij} be a set of exogenous covariates that might be of importance for modeling interactions. Our objective is to develop a statistical model of Y_{ij} by expressing p_{ij} as a function of these covariates. We model the dependency of p_{ij} on X_{ij} and Z_{ij} using the well-known *logistic regression models* in the statistical literature, i.e.,

$$\text{logit}(p_{ij}) \doteq \log \frac{p_{ij}}{1-p_{ij}} = \mu + \alpha^T Z_{ij} + \beta^T X_{ij}, \tag{3}$$

where μ, α, β are the vectors of unknown parameters that need to be estimated from the data, and ^{T} stands for transpose. Since X_{ij} are covariates related to the organization graph, we are particularly interested in quantifying β.

1) Model Fitting, Goodness of Fit and Significance Tests: Under the logistic regression model (3) and independence assumption, the log-likelihood of observed data is

$$\text{Log-likelihood} = \sum_{i,j} \left(Y_{ij} \log p_{ij} + (1-Y_{ij}) \log(1-p_{ij})\right), \tag{4}$$

where p_{ij} is given in (3), and the summation is taken over all pair of i, j. Let P be number of free parameters in the logistic regression model, and U be the total number of user pairs, then the degree of freedom of the logistic regression model is $L = U - P$. It is well known that when the regression model is the true model and when U is large, asymptotically, the deviance score, defined as

$$\text{Deviance} = -2\text{log Likelihood}, \tag{5}$$

where log Likelihood is defined in (4) is χ^2 distributed with L degrees of freedom. Furthermore, the difference in the deviance score in two nested models is also χ^2 distributed. Thus the deviance scores can be used for measuring goodness of fit of certain models, for model selection and for significance tests. We shall explain this later in more detail using concrete examples.

We use the well-developed iterative re-weighted least squares developed for generalized linear models to estimate the unknown parameters in the logistic regression model. In our study, we accomplish this using the *glm* routine in statistical language *R* [1].

2) Relation to Exponential Random Graph Models: Our statistical modeling methodology for user interaction graphs which uses logistic regression models is, in fact, closely related to the exponential random graph models (or p^* models) proposed in [2, 12]. In the framework of exponential random graphs, the joint distribution of linkage between nodes are modeled using local graph configurations. Dependence in linkages can be accommodated by considering complex local configurations such as two-star, three-star or triangle. Two methods have been developed to optimize the model parameters: Markov Monte Carlo maximum likelihood estimation and pseudo-likelihood estimation as an approximation technique. It has been shown that for large graphs, the two methods give estimates that do not differ significantly. Our logistic regression model is similar to the logistic regression approximations for the exponential random graphs under the simplifying assumption that edges are independent. The independence constraint can be relaxed by borrowing ideas from the framework of exponential random graph models.

We should comment here that our model has a much simpler form than the usual logistic regression approximation model in exponential random graph framework. To account for individual node level effects on interactions, the logistic regression approximation will have to use one covariate per node, thus creating a large set of covariates for a large network of many nodes. This increases the difficulty for model fitting significantly. In contrast, we eliminated this difficulty by using the observed node degrees as substitutes for node level activity, and treat the joint activity level as a single covariate in modeling the interactions.

B. Statistical Models

We first discuss the choice of covariates in the logistic regression model (3), and then present a few sets of models that incorporate different aspects of the organization graph.

1) Exogenous Covariates Z_{ij}: Each user has a different activity level that may impact his/her interactions. Furthermore, a user's level of interaction with others differs widely among users. It is important to include this effect in the model and differentiate it from the more interesting effects derived from organization graph.

Given the user population, for user i, let a_i be the total number of interactions with other users in the population, i.e., its degree in the interaction graph. Analysis of a_i for the interaction graph reveals that a_i follows a heavy-tailed distribution. For several heavy users, the number of interactions can reach up to 200, while around 90% users have less than 10 interactions.

Assume that a user interact with other users independently given his/her activity level a_i. As user population gets large, it is easily derived from random graph theory that the probability that the user pair (i, j) has an interaction is proportional to $a_i a_j$. In our statistical models, we shall include this pure chance activity effect as an exogenous covariate. Notice that the interaction graph is very sparse, and hence, p_{ij} are typically small; under the logistic regression model (3), we can use $\log(a_i a_j)$ as an exogenous covariate to represents their level of interaction by chance alone. Other exogenous covariates that might be important for modeling could be race and gender, but we do not consider them in this paper.

2) Covariates from Organization Graph X_{ij}: For a pair of users (i, j), we consider several covariates X_{ij} derived from the organization graph to characterize their relationship in the corporate hierarchy. For user i, let c_i be the country where he is from. For the pair of users (i, j) the indicator variable $I(c_i = c_j)$ is a suitable covariate representing whether or not users from the same country are more likely be linked. Let K be a set of possible countries, a more expansive form for characterizing how they are geographically alike are the set of indicator variables $I(c_i = k, c_j = l), k, l = 1, \ldots, K$ which identifies their country pair.

For a user pair (i, j), we extracted covariates related to their relative positions in the organization graph. The first candidate is the company hierarchy distance d_{ij} defined earlier as the number of hops to their nearest common ancestor, whichever is larger. For example, a distance 1 would imply either a boss/subordinate or a co-worker relationship. Since the hierarchical distance does not fully capture the relative position of the user pair in the organization graph, we supplemented this using additional covariates. For this purpose, we define a level-l organization as an organization of employees with a common ancestor at level l. For example, a level-1 organization includes all employees under the same person who is a direct subordinate under the CEO. Given that most employees are at level 5 in the hierarchy, for this study, the maximum level of organization we consider is 3. Therefore, we consider the set of indicator variables $Org_{ij}(l), l = 1, 2, 3$ as covariates.

3) Models in the Study: To limit the scale of the study (as the size of the vector Y_{ij} is $N(N-1)$ where N is the number of users), we eliminate Jive users with less than 5 activities in the 6-month period for this study. In addition, we only consider users from the top 10 countries. As a result, we have a total of 1284 Jive users, which results in a vector of Y_{ij} for all pairs of length 823686. Typical of a network graph, the value of Y_{ij} is mostly 0 except for about 4000 entries.

Let

$$r_{ij} = \log(a_i a_j), \tag{6}$$

where a_i, a_j are node degrees (number of interacting users) for i, j respectively. In the following, we list the four sets of statistical models that we use for analysis. First the basic model:

$$M_b : \text{logit}(p_{ij}) = \mu + \alpha r_{ij} \tag{7}$$

Recall we use c_i to represent the country of user i. The second set of models incorporate user geo-location (country) as covariates:

$$\begin{cases} M_{c1}: & \text{logit}(p_{ij}) = \mu + \alpha r_{ij} + \beta I(c_i = c_j), \\ M_{c2}: & \text{logit}(p_{ij}) = \mu + \alpha r_{ij} + \sum_{k,l=1}^{10} \beta_{kl} I(c_i = k, c_j = l), \end{cases} \tag{8}$$

where $\mu, \alpha, \beta, \beta_{kl}$ are the unknown parameters, where β representing the relative preference of a user pair from the same country to be linked together, and β_{kl} representing the relative preference of a user from a country k are linked to a user from a country l (also notice that 10 in the summations comes from the top 10 countries which is the scope of our models). It is easy to see the M_{c2} is an expanded model of M_{c1}, as it not only considers whether the users are from the same country, but also the distinct country pairs. It is interesting to study if adding this extra complexity is useful for predicting interactions.

Recall for a user pair (i, j), we use d_{ij} to represent their corporate hierarchy distance and $Org_{ij}(l), l = 1, 2, 3$ to present whether or not the user pair from the same level-l organization. The third set of statistic models incorporate covariates that characterize the relative positions in the corporate hierarchy. Ordered in increasing level of complexity, they are:

$$\begin{cases} M_{h1}: & \text{logit}(p_{ij}) = \mu + \alpha r_{ij} + \sum_{k=1}^{10} \gamma_k I(d_{ij} = k), \\ M_{h2}: & \text{logit}(p_{ij}) = \mu + \alpha r_{ij} + \sum_{k=1}^{10} \gamma_k I(d_{ij} = k) + \\ & \qquad \sum_{l=1}^{3} \eta_l Org_{ij}(l), \end{cases} \tag{9}$$

where $\mu, \alpha, \gamma_k, \eta_l$ are the unknown parameters, γ_k representing the relative preference of a user pair with hierarchy distance of k are linked together, and η_l representing the preference of a user pair being in the same level-l organization are linked together (note that the effect of $Org_{ij}(1), Org_{ij}(2)$ and $Org_{ij}(3)$ are nested).

Finally the full model that incorporates both covariates from user countries and their positions in the company hierarchy:

$$\begin{aligned} M_f: \quad \text{logit}(p_{ij}) = \mu + \alpha r_{ij} + + \sum_{k,l=1}^{10} \beta_{kl} I(c_i = k, c_j = l) \\ + \sum_{k=1}^{10} \gamma_k I(d_{ij} = k) + \sum_{l=1}^{3} \eta_l Org_{ij}(l). \end{aligned} \tag{10}$$

C. Results

1) Model Parameters: The fitted unknown parameters with respect to company hierarchy and geo-location is of great interest to us since they quantify the magnitude of effects. Interestingly, we have found that the fitted parameters of geo-location covariates and hierarchy level covariates remain relatively stable when we have either a more restrictive or expanded model. Table III and IV show the fitted parameters for two selective models, M_{c1} and M_{h1} respectively. They give quantification of the effects of "Same Country" and "Hierarchy Distance", as well as the exogenous variable such as the user pair activity level.

For both models M_{c1} and M_{h1}, the parameters of activity levels are very close to 1. This is, in fact, true for all our models. This implies that a random model between user pairs conditioning on their level of activity is a good base model.

Coefficients	**Estimate**	**Std. Error**	**z value**	$Pr(>|z|)$
Intercept	-9.75	0.06	-161.9	<2e-16
Activity level	1.03	0.011	93.55	<2e-16
Same.country	0.82	0.032	25.05	<2e-16

TABLE III
THE FITTING RESULT FOR MODEL M_{c1} IN (8). THE PREFERENCE BETWEEN USERS FROM DIFFERENT COUNTRY IS USED AS THE BASELINE. THE Z-VALUE IS CALCULATED AS ESTIMATE/STANDARD.ERROR, WHICH IS COMMENSURATE WITH A VALUE WITH GAUSSIAN DISTRIBUTION WITH VARIANCE 1.

Coefficients	**Estimate**	**Std. Error**	**z value**	$Pr(>|z|)$
Intercept	-5.4	0.10	-51.4	< 2e-16
Activity level	1.03	0.011	90.6	< 2e-16
Distance = 2	-1.35	0.12	-11.6	5.06e-06
Distance = 3	-3.18	0.11	-29.1	< 2e-16
Distance = 4	-3.96	0.10	-39.4	< 2e-16
Distance = 5	-4.30	0.10	-43.4	< 2e-16
Distance = 6	-4.44	0.11	-44.0	< 2e-16
Distance = 7	-4.39	0.10	-41.9	< 2e-16
Distance = 8	-4.49	0.15	-30.7	< 2e-16
Distance = 9	-3.79	0.51	-7.4	< 2e-16
Distance = 10	-4.01	0.51	-8.0	1.55e-15

TABLE IV
THE FITTING RESULT FOR MODEL (9). THE PREFERENCE BETWEEN HIERARCHY DISTANCE OF 1 IS USED AS BASELINE.

To interpret the effect of corporate hierarchy related covariates, we should first understand that model M_{c1} uses a different country as the baseline, and M_{h1} uses the hierarchy distance being 1 as the baseline. Since the fitted probability values p_{ij} is small compared to 1 and the model fits are done at the logit scale, Table III implies that if a pair of users are from the same country, then they are $e^{0.82} = 2.27$ times more likely to interact than if they are from different countries.

Similarly, Table IV indicates that users are more likely to interact if their hierarchy distance is small. For example, if the distance between a user pair is 2, than it is $e^{1.35} = 3.85$ times less likely to interact than if they are of distance 1 (which indicates a peering or boss/subordinate relationship). Similarly, if the distance between a user pair is 3, then it is $e^{3.18} = 24$ times less likely to interact than if they are of distance 1. However from the fit in Table IV, we also observe that the preference to interact does not exhibit significant differences when the hierarchy distance is equal to or larger than 5. The reason is that the company hierarchy is not a tree of equal depth in every branch. Many leaf nodes are distributed in level 5 and lower levels. As a result, many user pairs with a hierarchy distance equal to or larger than 5 are the users that are farthest away from each other in the company hierarchy, i.e., their nearest common ancestor in the company hierarchy tree being the root, the company CEO.

Our analysis suggests that users are more likely to interact with other users from the same country and closer in corporate hierarchy. This does not contradict to our earlier observations that the enterprise social network as a medium does bring users from diverse locations and social status together. This is because the number of users adjacent to a particular user in the organization graph is small as comparing to users that are far apart. So for users that are far part, even though the probability of their interactions is small, we still observe many occurrences of such interactions.

To provide further context for the interpretation, it should be noted that, except for the CEO and a few other highranking company employees that are well-known throughout the company, a given Jive user in an enterprise environment probably does not know the "rank" of another unfamiliar user outside of their organization. So although there already seems to be less interaction between users that are far apart in hierarchy, the fact that a given user may not know the rank of the person they are contemplating communicating with may have a "hidden" affect. The same is be true for the country information. So although the data suggests that users do communicate across country boundaries (albeit typically at a lower rate than within the same country), there may be some "hidden" affect on this since a given user may not know where another user is from.

It should be noted, however, that Jive users can determine the country of origin of another user rather easily by visiting the Jive "profile" page of that user. But users must consult a completely different corporate system if they wish to determine the position of another user within the corporate organizational structure, as it is not provided in the Jive interface.

2) Statistical Significance of Organization Graph Covariates and Model Comparison: We use the well-known statistical hypothesis testing procedure to test the overall significance of certain groups of covariates such as countries and hierarchy distances. Take the country models (8) as an example. To see if the addition of "country pairs" is significant on top of the effect of "same country", we test the following hypothesis:

$$H_0 : \beta_{k,l} = 0, \text{ for all } k, l = 1, \ldots, 10.$$

against

$$H_1 : \text{at least one } \beta_{k,l} \neq 0, \;\; k, l = 1, \ldots, 10.$$

To achieve this, let L_{c1}, L_{c2} be the degrees of freedom under the two models respectively (in this instance, $L_{c1} = U - 3, L_{c2} = U - 56$, where U is the total number of user pairs). Let D_{c1}, D_{c2} be the respective *deviance* scores of the two fitted models (5). Then we compare the value of $D_{c1} - D_{c2}$ with a χ^2 distribution with $L_{c2} - L_{c1}$ degrees of freedom. If the p-value associated with the χ^2 distribution is small, then we reject the null hypothesis, i.e., the effect is significant.

We fit all models (7), (8), (9), (10) using the *glm* (generalized linear model) routine in the statistical software R. Table V lists the deviance of the fitted models with their respective degrees of freedom. A graphical representation is also shown in Figure 4, where different colors represent different sets of models as defined earlier.

Model	Covariates	Degrees of Freedom	Deviance
M_b	Basic	$U-2$	41069
M_{c1}	SameCountry	$U-3$	40477
M_{c2}	CountryPair	$U-56$	39799
M_{h1}	HierDist	$U-11$	38677
M_{h2}	HierDist+SameOrg	$U-14$	37733
M_f	All	$U-68$	36893
M_{com}	Community	$U-50$	39130
M_{f+com}	Community + All	$U-116$	35194

TABLE V

DEVIANCE SCORE ($-2\log$ LIKELIHOOD) OF THE FITTED MODELS FOR THE USER INTERACTION GRAPH. $U = 823686$ IS THE TOTAL NUMBER OF USER PAIRS.

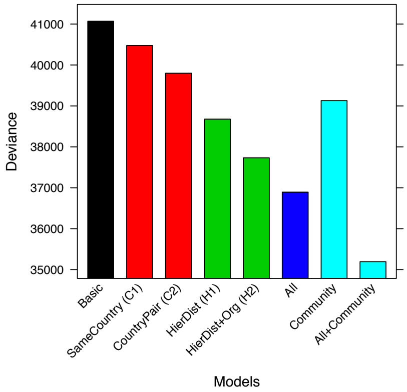

Fig. 4. Deviance (-2log Likelihood) of the 8 statistical models in (7), (8), (9), (10) and (11). Colors indicate the distinct sets of models.

To see whether the geographic locations and corporate hierarchy covariates are significant, we use the χ^2 significance test that we described earlier. Table VI shows the significance of covariates under consideration. As we can see from Table VI, all organization graph-related effects are highly significant. In particular, the addition of hierarchical level covariates yields a bigger reduction in the deviance score than the geo-location covariates, indicating the more important role of corporate hierarchy in predicting the user interactions. Also interestingly, the SameOrg provides significant additional improvement over models using hierarchical distances alone.

Covariates	Models	Deviance Diff	Reference χ^2 99% quantile
SameCountry	M_b vs. M_{c1}	592	χ^2(1)=6.6
CountryPair-SameCountry	M_{c2} vs. M_{c1}	678	χ^2(53)=79.8
CountryPair	M_{c2} vs. M_b	1270	χ^2(54)=81
HierDist	M_{h1} vs. M_b	2392	χ^2(9)=21.7
SameOrg	M_{h2} vs. M_{h1}	944	χ^2(3)=11.3

TABLE VI

SIGNIFICANCE TESTS FOR COVARIATES FROM THE ORGANIZATION GRAPH. THE INTEGER IN $\chi^2(\cdot)$ INDICATES THE DEGREE OF FREEDOM OF THE REFERENCE χ^2 DISTRIBUTION.

D. Communities as Additional Covariates

As the enterprise social network medium bring users from diverse locations and social status forming ad-hoc communities, as a further improvement, we considered including the user communities as potential covariates for modeling the user interactions. We used communities discovered using the leading eigenvectors of the adjacency matrix discussed in Section IV-B for this purpose.

Let g_i be the community of user i. We consider the following two models.

$$\begin{cases} M_{com} : \text{logit}(p_{ij}) = \mu + \alpha r_{ij} + \\ \qquad \sum_{k,l=1}^{7} \phi_{kl} I(g_i = k, g_j = l) \\ M_{f+com} : \text{logit}(p_{ij}) = \mu + \alpha r_{ij} + \\ \quad \sum_{k,l=1}^{10} \beta_{kl} I(c_i = k, c_j = l) + \sum_k \gamma_k I(d_{ij} = k) + \\ \quad \sum_{l=1} \eta_l Org_{ij}(l) + \sum_{k,l=1}^{7} \phi_{kl} I(g_i = k, g_j = l), \end{cases} \tag{11}$$

where M_{com} is the model with community covariates alone, and M_{f+com} is the complete model with both corporate hierarchy related and community covariates.

Table V as well as Figure 4 shows the resulting fit of two models. From the result, it is obvious that the effect of the community is significant and provides additional improvement to the overall fit. Comparing M_f and M_{f+com}, the difference in deviance is 1699, while the reference value of a χ^2 distribution with 48 degrees of freedom is only 73.7. It is also interesting to observe that the community covariates alone is not sufficient to substitute the corporate hierarchy related effects, since M_{com} has a much higher deviance score than M_f. In fact, we observe that the effect introduced by communities is almost orthogonal to that of organization graph covariates, and the fitted parameters for organization graph covariates are almost unchanged when including the community covariates.

E. Summary of Results

In this section, we built formal statistical models to quantify the effect of user's attributes from the organization graph on their interaction patterns in the enterprise social network. Our models are based on logistic regression, and are related to but not identical to the exponential random graph models (or p^* models) proposed in [2, 12]. Through analysis, we have found both user's geo-location and position in corporate hierarchy are highly significant in predicting their interactions. For example, if a pair of users are from the same country, then they are 2.27 times more likely to interact than if they are from different countries. Furthermore, users are more likely to interact if their hierarchy distance is small. As an another example, if the hierarchical distance between a user pair is 2, than it is 3.85 times less likely to interact than if they are of distance 1 (which indicates a peering or boss/subordinate relationship). Finally, as the enterprise social network medium brings together users from diverse locations and social status forming ad-hoc communities, we also discovered that including these communities in the statistic models improves the fit significantly.

VI. Related Work

We divided the related work into two categories:

Enterprise social networks: There are several studies on research prototypes of enterprise social networks. Brzozowski introduces a social media aggregator named WaterCooler in HP and studies the users' behavior [3]. He uses case studies to show that geographically dispersed teams are more prone to use enterprise social networking applications. Kolari *et al*. analyze the graph structure and properties of the use of an internal corporate blog service [9]. They also study the overall distribution of interactions across different hierarchy distance. In comparison, the enterprise social network we study provide blog service as well as three other types of services, and we show that the characteristics of blog service is significantly different from discussion and microblog. Our study yields consistent results, and further reveals how the hierarchy distance affects the interaction between distinct user pairs.

Analysis of public social networks: Gupte et al [7] try to infer social hierarchy from social networks. They show that hierarchy emerges as the size of the network increases. Further, they show that the degree of stratification in a network increases very slowly as the size of the graph increases. In our enterprise social network setting, hierarchy is known. Xiang et al [15] develop an unsupervised model to estimate relationship strength from interaction activity (*e.g.* communication, tagging) and user similarity. In contrast, we specifically focus on the relationships that exists in enterprise social networks. These relationship such as supervisor-subordinate, co-worker, *etc.* , exhibit interesting user behavior.

VII. Conclusions

In this paper, we take the first step in analyzing and modeling user interaction in enterprise social networks. User interaction is a tale of two graphs: the organization graph and the social interaction graph. We build a formal model to explain such interactions, and demonstrate that two user attributes, user geo-location and position in corporate hierarchy are highly significant in predicting user interactions.

References

[1] The R Project for Statistical Computing. http://www.r-project.org.

[2] Anderson, C., Wasserman, S., and Crouch, B. A p* primer: logit models for social networks. *Social Networks* (1999).

[3] Brzozowski, M. J. Watercooler: exploring an organization through enterprise social media. In *Proceedings of the ACM 2009 international conference on Supporting group work* (New York, NY, USA, 2009), GROUP '09, ACM, pp. 219–228.

[4] Clauset, A., Newman, M. E. J., and Moore, C. Finding community structure in very large networks. *Physical Review E 70*, 6 (Dec. 2004), 066111+.

[5] Dion Hinchcliffe. Social business and enterprise usage: The lessons. http://www.zdnet.com/blog/hinchcliffe/social-business-and-enterprise-us%age-the-lessons/1882?tag=content;siu-container, December 2011.

[6] Dion Hinchcliffe. Enterprise 2.0 Success: Alcatel-Lucent. http://www.zdnet.com/blog/hinchcliffe/enterprise-20-success-alcatel-luc%ent/1917?tag=content;siu-container, Jan 2012.

[7] Gupte, M., Shankar, P., Li, J., Muthukrishnan, S., and Iftode, L. Finding hierarchy in directed online social networks. In *Proceedings of the 20th international conference on World wide web* (New York, NY, USA, 2011), WWW '11, ACM, pp. 557–566.

[8] Jive Software Inc. Jive Social Business - Collaboration & Social Software Solutions. www.jivesoftware.com/.

[9] Kolari, P., Finin, T., Yesha, Y., Yesha, Y., Lyons, K., Perelgut, S., and Hawkins, J. On the Structure, Properties and Utility of Internal Corporate Blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)* (March 2007). Nominated for Best Paper Award.

[10] Lauren Fisher. 44% of companies track employees' social media use in and out of the office. "http://thenextweb.com/socialmedia/2011/08/17/44-of-companies-track-emp%loyees-social-media-use-in-and-out-of-the-office/, August 2011.

[11] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. Measurement and analysis of online social networks. In *Proc. of ACM SIGCOMM Internet Measurement Conference* (2007).

[12] Robins, G., Pattison, P., Kalish, Y., and Lusher, D. An introduction to exponential random graph (p) models for social networks. *Social Networks 29*, 2 (2007), 173–191.

[13] Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P., and Zhao, B. Y. User interactions in social networks and their implications. In *Proc. of EuroSys* (2009).

[14] www.informationweek.com. Social Networking: Set Internal Collaboration Goals Early. http://www.informationweek.com/thebrainyard/news/social_networking_priv%ate_platforms/231900531/social-networking-set-internal-collaboration-goals-ear%ly, Oct 2011.

[15] Xiang, R., Neville, J., and Rogati, M. Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on World wide web* (New York, NY, USA, 2010), WWW '10, ACM, pp. 981–990.