Communication Structure Discovery via Information Asymmetry in an Organizational Social Network

Cheng-Te Li and Shou-De Lin
Graduate Institute of Networking and Multimedia, National Taiwan University
{d98944005, sdlin}@csie.ntu.edu.tw

Abstract

In an organization, based on the positions of employees there is usually an existing hierarchy among them. However, in real-life cases, people's interactions tend to form a certain communication structure due to some external forces or personal factors. In this paper, we aim at discovering the potential communication structure, in which nodes are typed labels (e.g. job-titles) and edges stand for tight interactions between typed labels in an organizational social network. To tackle this problem, we propose to exploit the concept of information asymmetry to model the core-periphery property in the communication structure. The proximity asymmetry is defined to realize the information asymmetry. We also devise two random-walk methods to calculate the proximity asymmetry between typed labels. The experiments conducted on the Enron email dataset shows that the proposed method outperforms some heuristic ones.

1. Introduction

The communication structure is a kind of graph structure that captures the potential core-periphery interaction patterns among different types of entities in an organizational social network. The communication structure describes which typed individuals frequently interacts with one another and how the information flow among typed individuals along the network. The communication structure is different from the formal hierarchy [4] of titles in that the former depends on how the information spreads across typed entities in the network while the latter is established rigidly based on the existing positional hierarchy defined by the highest level of management in the cooperation. Examples of both cases are given in Figure 1.

For the purpose of the effective and efficient management, most organizations would ask their members to issue commands or diffuse information given certain standard procedures. The information usually flows along formal hierarchy in a top-down manner. For example, the hierarchy of a corporation usually consists of CEOs, presidents, directors, managers, staffs, and etc. However, for different reasons, the real-life scenarios of communications might not follow this case. People who were lower in the formal hierarchy might play more significant roles in delivering orders while some people with certain titles might seldom communicate with the subordinates or hardly get involved in communication. Therefore, in this paper, we aim at designing a method that is capable of discovering the true communication structure in an organizational social network.

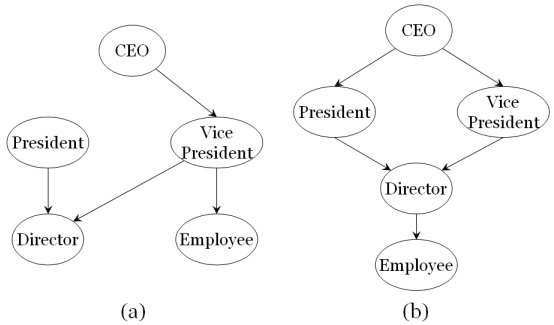

Figure 1: For job-titles, (a) the potential communication structure, (b) the illustrated formal hierarchy.

For the hierarchy of typed labels, G. M. Namata et al. [2] proposed a classification-based formal hierarchy inference for titles. Their assumption is that two titles with direct parent-child pairs will have overlapped works and behave similarly to some extent. Thus, those misclassified titles are used to construct the hierarchy. However, this method needs some training data as well as certain content information to generate similarity pairs. Our method is an unsupervised one and do not need extra information. Later, they [3] assume those with the same title would interact closely, and thus propose a clustering-based method for the hierarchy of titles. Nevertheless, this assumption does not always hold for all kinds of network. Those individuals of the same typed label are not necessary to form tight structures.

This paper proposes a method to discover the communication structure of typed labels in an organizational social network. The central idea is that the information flows between the core-to-periphery and periphery-to-core directions in the social network are asymmetric. We propose to exploit the proximity to estimate the information propagation across typed labels. Two random-walk-based mechanisms, including separated and grouped aspects, are devised to compute the pairwise proximity. And then we derive a ranked list of typed pairs by maximizing the proximity asymmetry. Finally, we present a greedy method to construct the communication structure by the ranked list.

2. Methodology

Definition 1 (Organizational Social Network). A organizational social network $G = <V,E,L>$ is a undirected graph, where V is a finite set of vertices, $E \subseteq V \times V$ is a finite set of edges, and L is also a finite set of typed labels. The function $types(V) \rightarrow \{\{t_1,\ldots,t_k\}, t_i \in L, k \geq 1\}$ maps each vertex into a type label. The label associated with each node corresponds to its categorical attribute (e.g., the job-title: "CEO", "manager", "employee", etc.).

978-0-7695-4191-4/10 $26.00 © 2010 IEEE
DOI 10.1109/WI-IAT.2010.171

Definition 2 (Communication Structure). A communication structure $H=<V_H,E_H>$ is a connected directed acyclic graph, where $V_H=\{t_1,\ldots,t_k\}$ ($t_i \in L$, $k \geq 1$) is a finite set of vertices that corresponds to typed labels, and E_H is a finite set of directed edges. Note that there are disjoint sets of nodes for each typed label t_i, $S_{t1}=\{v_{11},\ldots,v_{1n}\},\ldots,$ $S_{tk}=\{v_{k1},\ldots,v_{kn}\}$, $S_{ti} \cap S_{tj}=\phi$, $i \neq j$.

We aim at discovering the communication structure. Each node in the structure is a typed label. Each directed edge e_{ij} reflects the relationship from a certain leader type i to a certain follower type j. The meaning of the relationship from i to j could be "interact closely with", and "distribute information to". The key of our approach is the information asymmetry, stating the imbalance of information diffusion between the core-to-periphery and the periphery-to-core pairs of individuals.

2.1 Information Asymmetry

In economics, the common definition of information asymmetry refers to a situation in which at least certain information is known to some people but not all individuals of that event. That is, there exists one party in a transaction which has superior information than another [1]. For example, sellers are usually inclined to know more about the commodities than buyers. We take advantage of this concept to find the communication structure. The hypothesis is that individuals in core positions of the communication structure would have higher chance to send information to the peripheral ones while theose in peripheral positions are relatively rare to do so. In other words, there is a kind of asymmetry or imbalance of information between two parties. Hence, we define and measure the asymmetry between two parties. Note the parties are referred to different typed labels on nodes. We propose the pairwise proximity as the asymmetry measure.

2.2 Proximity Asymmetry

To measure the extent of proximity, we compute the volumes of information communication between any two typed labels t_i and t_j in the network. That is, the proximities of communications from t_i to t_j and from t_j to t_i are calculated. We provide an example shown in Figure 2, to illustrate the idea of asymmetry of proximity between central-outer pairs. Figure 2(a) is a communication network. There are five typed labels. First, the communications between superior-to-inferior pairs are not so frequent. Some interactions, such as CEO-to-VicePresident and VicePresident-to-Employee, are more frequent. Second, the interactions for the inversed directions of the above pairs are relatively rare. Besides, some communications between pairs of indirect inferior-to-superior typed labels are infrequent as well, such as Director-to-CEO and Employee-to-President.

We summarize the proximity of these interactions between typed labels by Figure 2(b) and the resulted communication structure is shown in Figure 1(a). The directed bold lines indicate the high proximity of communications from central typed labels to outers. The directed dotted lines show the low proximity of interactions from a) a certain central ones to its indirect peripheral ones, b) a certain peripheral ones to its immediate centrals, and c) a certain peripheral to its indirect centrals. For example, the vice presidents have high proximities to directors and employees while they have low ones to CEO, and presidents.

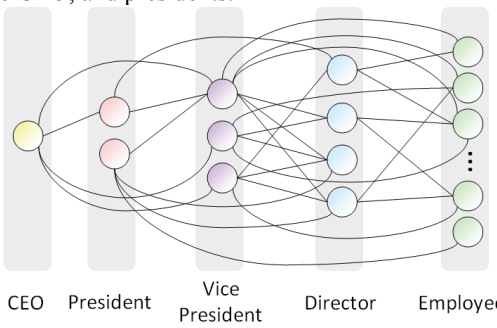

Figure 2(a): An illustrated network for different titles.

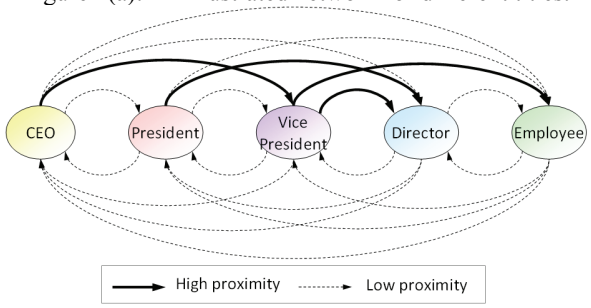

Figure 2(b): An illustration of command flows for the titles.

Using this idea, we can calculate the pairwise proximity between typed labels to estimate the asymmetry. We first formally define the pairwise proximity asymmetry.

Definition 3 (Pairwise Proximity Asymmetry). Given a *matrix* of pairwise proximity M_{PP}, in which each entry d_{ij} is the pairwise proximity value between a group of nodes associated with typed label t_i and the other with t_j, the pairwise proximity asymmetry between t_i and t_j is defined as $M_A(i,j) = |M_{pp}(i,j) - M_{pp}(j,i)|$. Note the diagonal of the matrix is set to 0. And the M_{PP} is an asymmetric matrix and the M_A is a symmetric matrix.

If we pick a certain typed pair (t_i, t_j) in core position of the structure, the difference of pairwise proximities from two directions is large. Namely, to explore the communication structure, we have to maximize the pairwise proximity asymmetry $M_A(i,j)$ among all typed pairs (t_i, t_j). We sort the $M_A(i,j)$ in an ascending order, and a ranked list of pair of typed labels can be derived. Those pairs in top positions will have higher potentials to be the edges of the resulted structure. This list is called Ranked Pair List of Pairwise Proximity, denoted by RPL_{PP}. The procedure to obtain the RPL_{PP} is formulated by the following equation.

$$RPL_{PP} = \arg\underset{(ti,tj)}{\left(sort\left(\underset{ti,tj \in L, i \neq j}{\forall} M_A(i,j) \right) \right)}, s.t. M_A(i_1,j_1) > M_A(i_2,j_2)$$

2.3 Proximity Computation

We exploit the mechanism of *random walk with restart* (RWR) [6] to calculate the proximity information between a certain type t_i and the other t_j. The main idea of RWR is to use the random suffer mechanism to propagate information from an indicated source node. Then the affinity scores of surrounding nodes with prior to the source one can be computed based on the structural proximity. We divide the calculation of RWR proximity $Prox(S_{ti}, S_{tj})$ for a node set with typed labels t_i to the other of t_j into two aspects. One is from the viewpoint of each pair of separated nodes of both indicated sets called *separated random walk* while the other is from the view of pairs of indicated groups called *grouped random walk*.

2.3.1 Separated Random Walk. To attain the proximity between two typed sets S_{ti} and S_{tj}, we compute the proximities for all pairs of individuals of two different typed sets and combine the separated proximity score to be the proximity of a pair of typed set. In detail, taking each node i from S_{ti} as the source, the random walker iteratively transmits to its neighborhood with a uniform probability that is inversely proportional to its degree, and also at each step it has some probability c to return to the source node i.

Here we give the method to compute the separated proximity. Let M_{ti} is a $N\times|S_{ti}|$ matrix, in which each column corresponds to a node of t_i and all its entries are zero, except for the one entry that corresponds to node $i \in S_{ti}$ (set this entry to 1). N is the number of nodes and A is the adjacency matrix of the graph G, which is column-normalized. $M_{prox,ti}$ is the $N\times|S_{ti}|$ proximity matrix, which will be iteratively computed to derive the steady-state probabilities from each node of t_i to others in the end. Besides, c is the restart probability. Then we can derive $M_{prox,ti} = (1\text{-}c)AM_{prox,ti} + cM_{ti}$. We can further put all $t_i \in L$ (L is a finite set of typed labels of graph G) together to form a large matrix M_{prox} so that we can efficiently derive the proximities from nodes of each type using one iterative matrix multiplication. Note M_{prox} is a $N\times N$ matrix. The same aggregation is performed on M_{ti} to M_s. It is shown in the following equation, where l is $|L|$.

$$\begin{aligned} M_{prox} &= [M_{prox,t1}, M_{prox,t2}, ..., M_{prox,tl}] \\ &= (1-c)A[M_{prox,t1}, M_{prox,t2}, ..., M_{prox,tl}] + c[M_{t1}, M_{t2}, ..., M_{tl}] \\ &= (1-c)AM_{prox} + cM_s \end{aligned}$$

To attain the separated pairwise proximity matrix M_{PP}^{sep} by separated random walk for information asymmetry, where each entry (i, j) in M_{PP} is the proximity from t_i to t_j, we sum up all proximities from the instances of a certain type to another. It can be determined by $M_{pp}^{sep}(t_i,t_j)=(\Sigma_{v\in Stj}\Sigma_{u\in Sti}M_{prox}(v,u))/|S_{ti}|$, where $t_i \neq t_j$. Therefore, if we want to compute the separated pairwise proximity $M_{PP}^{sep}(t_i, t_j)$, those column vectors with typed label t_i are picked, and those entries with typed label t_j in each column vectors are sum up. Besides, we normalize the summed value by the number of nodes of t_i to derive the mean proximity of a certain node of t_i to t_j.

2.3.2 Grouped Random Walk. Instead of concerning about the volume of information flows for each individual of a certain type behaves (i.e., the average proximity from one type to another), the grouped approach focuses on the integral information accessibility from one type to another. Thus, we create a supernode to gather each node of a certain type t_i together. By seeing them as a unity, the random walk with restart is employed again to compute the proximity from the supernode of t_i to the other types. As for those edges originally connected to each individual of t_i, they simply count once if a certain node of $t_j \neq t_i$ has links to multiple individuals of t_i. Since both A and C belong to t_i, there is only one edge between node I and the supernode of t_i. We call the graph with a t_i supernode the type-grouped graph G_T. Thus, for each t_i in L, we have a $G_{T,ti}$. And for each $G_{T,ti}$, there is only one node for t_i. Eventually, the grouped pairwise proximity matrix M_{PP}^{grp} is computed using the type-grouped graph $G_{T,ti}$ by applying the separated one to derive the proximity from t_i to other types.

2.4 Communication Structure Construction

As aforementioned, we can generate the *Ranked Pair List* (RPL) that estimate the pairwise proximity asymmetry by the computed separated and grouped pairwise proximity M_{PP}^{sep} and M_{PP}^{grp} for all typed pairs in the network. Since those in central positions of RPL hold higher imbalance scores of communications, they tend to form links in the communication structure. Therefore, based on the RPL and a given threshold k to pick pairs of high asymmetry scores, we devise a greedy method to construct the structure of typed labels. In the greedy construction, if a pair of typed label t_i and t_j is picked, two nodes are added for t_i and t_j, and a directed links e^H_{ij} is added to the final structure. The direction is determined by the pairwise proximities. If the pairwise proximity from t_i to t_j is larger than that of from t_j to t_i, t_i is regarded as the leading type while t_j is the following one. The direction of e^H_{ij} is drawn from t_i to t_j. Besides, in the method, we do not allow cycles in the final structure because the cyclic graph will cause a certain contradiction for the core-periphery intents. It is constructed by iteratively adding the top k pairs.

3. Evaluation

We use the Enron email dataset [5] in our evaluation. In the dataset, there are 151 employees and each has a formal job title. To construct the organizational social network, let the N_{AB} as the number of A sends a mail to B and N_{BA} for B sends to A. If $N_{AB} > 1$ and $N_{BA} > 1$, we construct an edge from A to B. Note we consider the "send" and ignore the "reply" messages since we focus on the original information or command flows. The constructed network contains 151 nodes and 516 edges. Taking the titles associated with each individual as typed labels on nodes, our method automatically constructs the communication structure of job titles from the

network. Note this data is incomplete due to 46 of 151 listed as *N/A*. The numbers of individuals for titles are CEO: 4, President: 4, Director: 13, Trader: 12, Managing Director: 3, Vice President: 19, Manager: 10, Employee: 37, Lawyer: 3, and N/A: 46. We manually find a communication structure from the Enron email dataset, and regard this structure as the ground truth containing 9 typed nodes and 16 directed links, as shown in Figure 3. By comparing the top-k returned pairs and the ground truth, we calculate the precision and recall scores to demonstrate the effectiveness of proposed method.

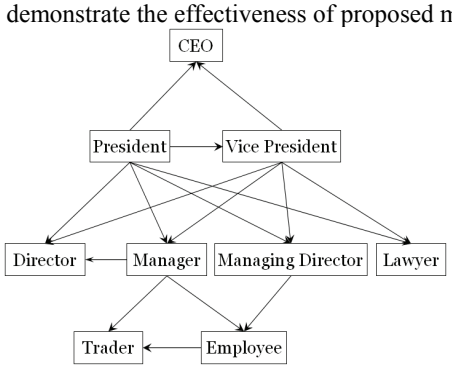

Figure 3: The observed Enron communication structure.

We compare our random-walk-based methods with three heuristic methods for the pairwise proximity between typed labels. Note not only the random walkers but also these heuristic methods follow the proposed idea of information asymmetry. The first heuristic is "mutual sent" (*MutualSent*). It measures the pairwise proximity by counting the mails sending from one to the other. The second is "minimum shortest distance" (*MinShortDist*). It uses the average of the minimum shortest distances from individuals of t_i to that of t_j as the proximity. The idea is the interactions between two types might try to find the convenient ways to transmit messages. The third is "frequent shortest distance" (*FreqShortDist*). For each individual of t_i, it finds the frequent shortest distance to that of t_j, and use these distances as the proximity. The intuition is that the regular communication behaviors among individuals are regarded as the proximity.

The result is reported in Figure 4 and Figure 5. The random walks averagely outperform the three heuristic methods significantly. We believe it is because the random walkers keep some key factors at one time: 1) the length of connections, 2) the quality of intermediated nodes, 3) the multiple connections, and 4) the neighbored structure of the source nodes. The three heuristics might ignore some of these factors when measuring proximities. Besides, we can observe the grouped random walk outperform the separated one. The precision of the grouped random walker reaches 0.867 and its recall is 0.813 when top 15 pairs are returned. This result shows the effectiveness of integral information for finding the communication structure. We think it is because: (1) the grouped one coordinates the proximities of individuals of the source type simultaneously while the separated one neglects the potential overlapped proximities to others for the individuals of the source type. (2) The hidden relationships among individuals of the source typed label should be considered to capture the integral authority of communication in an organization.

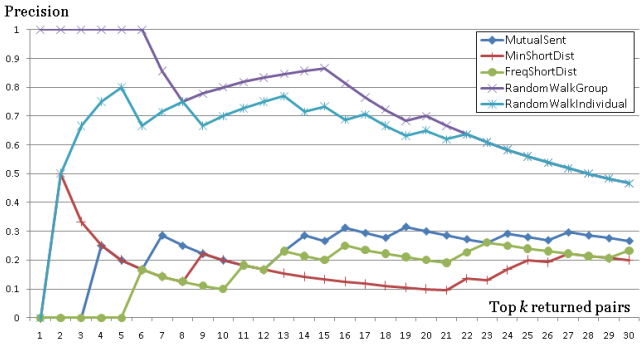

Figure 4: Precision curves w.r.t. the top k returned pairs.

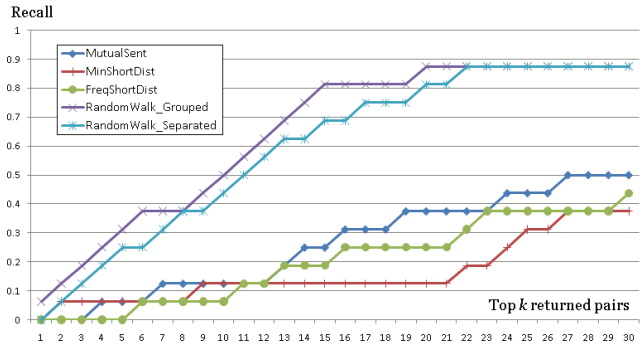

Figure 5: Recall curves w.r.t. the top k returned pairs.

4. Conclusions

We find the communication structure of typed labels from an organization social network. The central idea is information asymmetry stating that the proximity of the core-to-periphery directions tends to be higher than that of the periphery-to-core ones in the structure. We propose the separated and grouped random walks to compute the proximity between pairs of typed labels. By maximizing the proximity asymmetry and deriving the ranking list of typed pairs, we devise a greedy method to construct the communication structure. The evaluation shows the proposed approach outperform some heuristic ones.

References

[1] G. A. Akerlof, The Market for Lemons: Quality Uncertainty and the Market Mechanism, *Quarterly Journal of Economics*, 84(3), 488-500, 1970.

[2] G. M. Namata, L. Getoor, and C. Diehl, Inferring Formal Titles in Organizational Email Archives, *ICML Workshop on Statistical Network Analysis*, 2006.

[3] G. M. Namata, B. Staats, L. Getoor, and B. Shneiderman, A Dual-View Approach to Interactive Network Visualization, *ACM Intl. Conference on Information and Knowledge Management (CIKM'07)*, 939-942, 2007.

[4] E. Ravasz and A. L. Barabasi, Hierarchical Organization in Complex Networks, *Physical Review E*, 67, 026112, 2003.

[5] J. Shetty and J. Adibi, The Enron Email Dataset Database Schema and Brief Statistical Report, *Technical Report, ISI*, 2004.

[6] H. Tong and C. Faloutsos. Center-piece Subgraph: Problem Definition and Fast Solution. *ACM Intl. Conference on Knowledge Discovery and Data Mining (KDD'06)*, 404–413, 2006.