SSRM: Structural Social Role Mining for Dynamic Social Networks

Afra Abnar, Mansoureh Takaffoli, Reihaneh Rabbany, Osmar R. Zaïane
Department of Computing Science, University of Alberta,
Edmonton, Alberta, Canada T6G 2E8
Email:{aabnar,takaffol,rabbanyk,zaiane}@ualberta.ca

***Abstract*—A social role is a special position an individual possesses within a network, which indicates his or her behaviours, expectations, and responsibilities. Identifying the roles that individuals play in a social network has various direct applications, such as detecting influential members, trustworthy people, idea innovators, etc. Roles can also be used for further analyses of the network, e.g. community detection, temporal event prediction, and summarization. In this paper, we propose a structural social role mining framework (SSRM), which is built to identify roles, study their changes, and analyze their impacts on the underlying social network. We define fundamental roles in a social network (namely *leader*, *outermost*, *mediator*, and *outsider*), and then propose methodologies to identify them, and track their changes. To identify these roles, we leverage the traditional social network analyses and metrics, as well as proposing new measures, including community-based variants for the Betweenness centrality. Our results indicate how the changes in the structural roles, in combination with the changes in the community structure of a network, can provide additional clues into the dynamics of networks.**

INTRODUCTION

Hundreds of years ago, people traveled in form of caravans through the Silk Road for thousands of miles. These caravans connected cultures and economies from China to Mediterranean regions and Europe. What makes them interesting in our contexts, is how the responsibility of travelers in these caravans were correlated with their actual position in the caravan. For example, the individuals who knew the path and were leading the caravan would have been located in the front section, whereas warriors who were responsible to protect the caravan would have been located in both front and back sections, as well as all along the caravan to keep everyone and especially the merchandise safe.

Similar *organized structure* exists in most other human societies, where the position of an individual within the society correlates with his or her social role. The social *role* is in fact an important and well-studied concept in sociology. This concept refers to a special position, which is associated with a set of behaviours, expectations, and responsibilities [1]. People play different roles in their communities and societies, behave in special ways, and abide by certain expectations. Social roles are mainly used to define the influence of members on the network structure. As Walton [2] states, the changing nature of an individual and its leadership are of central importance to the explanation of social action.

In social network analysis, various metrics are defined for quantifying structural and non-structural (behavioural) characteristics of individuals, such as different centrality measures [3]. However, there exists a gap between these metrics and a high-level definition for the roles of individuals. In this paper, we propose the "Structural Social Role Mining" (SSRM) framework to define and identify these high-level roles. This framework unifies various kinds of metrics including the well-known centrality measures [3], and also different social network analyses such as community detection, influence propagation, and diffusion of innovation into the high-level concept of social roles.

Our proposed social roles provide more insightful information compared to their underlying social network metrics. In a parallel work under review, we have shown how these social roles can be effectively used to predict the temporal events and transitions taking place in dynamic social networks. Moreover, classifying nodes based on these social roles can provide another level of abstraction for the large social networks; thus, leading to more scalable and efficient algorithms. The three main contributions of this paper can be highlighted as:

1) A methodical framework for defining, and identifying structural roles is proposed;
2) Role changes are tracked over time and their correlation with the structural events in the network is illustrated;
3) New metrics, based on the community structure, are proposed, which are necessary for detecting social roles.

In the rest of the paper, after overviewing the related work, we define a set of fundamental roles in a network, which are based on the structural properties of the individuals, as well as the structure of the whole network. Then, we present metrics and methodologies to identify the aforementioned roles, and to study changes of roles in consecutive time frames. We define *role events* that capture the temporal changes of roles, and study the relation between these events and other structural events in the network. In the experiments, we show how exploring these roles and their impacts gives valuable insight into the evolution of the network.

RELATED WORK

Role is an important concept in sociology, however, there is no consensus on the definition of role between sociologist. Biddle [1] integrates various theories on role and discusses about functional, structural, organizational, and cognitive role theories. Among different theories on role, the only one which enables modeling the concept mathematically is the structural role theory. Oeser et al. [4]–[6] develop a mathematical model for structural role theory by defining three component *task*,

position, and *person* that define a role in connection with each other and also with other positions in a society.

The emergence of social networking tools enables access to more information in order to model and study social roles. Consequently, the study of roles is now becoming an interdisciplinary field of research, attracting researchers from different disciplines, specifically data mining and machine learning. Forestier et al. [7] present a survey of the state-of-the-art techniques for role mining in social networks. They further categorize roles to explicit, and non-explicit. *Explicit roles* are defined a-priori, and are identified by calculating a specifically designed method or a predefined criteria. Whereas, *non-explicit roles* are identified in an unsupervised framework, which requires little information about the roles beforehand. Clustering algorithms are usually used to identify non-explicit roles based on structural or contextual information in a network.

The two widely defined examples of explicit roles are *experts* and *influentials*. Notably, Zhang et al. [8] identify the expert role on a java technical forum. They propose three algorithms based on z-score, pagerank [9] and HITS [10] using both indegree and outdegree of nodes to identify experts. When they compare their results to the results on simulated networks, they observe that *the structure of the network has a significant impact on the ranking of experts*. Identifying nodes as *influential* role has also attracted considerable attention from researchers, mostly due to the influential role's wide range of applications in viral marketing, and diffusion of information. Kim and Han [11] distinguish three types of influential roles: *sales person*, *opinion leaders*, and *connector*. They further develop a two-step methodology for identifying the first type based on structural properties of the network. Agarwal et al. [12] explicitly define influential as an individual who is prominent in diffusion of innovation. They identify influential bloggers in blogosphere by defining an *iIndex* for each blogger based on their influential blog posts. Influential bloggers are those who have at least one influential post. In addition, authors discuss that influentials are different from initiators of an idea or creators of a content. *Influentials are more important because of their position in the network that empowers them to diffuse the influence*.

In all the aforementioned works, the community structure of the network is not directly considered in identifying roles. Community-based roles, on the other hand, are less studied, while they are important in many contexts, including link-based classification and influence maximization, as shown in [13]. Scripps et al. [13] define four structural community-based roles (*ambassadors*, *big fish*, *loners*, and *bridges*); and identify them based on the degree of nodes, and their community affiliation. Ambassadors are defined as nodes with high degree and also high community metric. Whereas a big fish is an individual who is only important within his/her community. Bridges are the individuals with high community score, but low degree, and loners are the ones with low degree and also low community metric. These roles are defined statically, and depend strictly on their metrics. However, the framework presented in this paper delivers a more generalized methodology for identifying roles, and could be used with various metrics, while also tracks the temporal changes of roles.

STRUCTURAL SOCIAL ROLE DEFINITION

Societies can be studied though social networks. A social network is modelled by a graph $G(V, E)$, where V is the set of vertices/nodes and E is the set of graph's edges. Here, the entities/individuals in the society are associated with the nodes of the graph, whereas the connections/interactions between them are represented with the edges. Characteristics and attributes of entities and their interactions can be also included in this model, as different attributes on the nodes and edges, which depend on the context.

We categorize the information in a social network into *structural* and *non-structural* properties. Structural properties are related to the topology of the graph such as an entity's connections (edges), neighbourhood structure, and the entity's position in that structure. Whereas, non-structural properties are the information not reflected in the topology of the graph, such as entities' attributes, connections' attributes and meta-data about the graph. In a static network, role of an entity is determined by its structural and non-structural properties. These properties may change in the temporal networks. Thus, we further define *temporal-structural* and *temporal-non-structural* properties to reflect the changes in the features of the network over the observation time. Since roles are important in the way that they affect their environment, these temporal properties should also be considered in identifying roles in a temporal network.

Our proposed SSRM framework is built upon two characteristics of human societies. The first is the role-taking behaviour of the individuals in interaction with each others which is aligned with their structural properties. The second is the fact that all societies are intrinsically modular i.e. composed of multiple groups (a.k.a communities). Thus, we define roles of individuals in a social network considering only structural properties of nodes, i.e. taking into account their their interactions with other individuals, along with their affiliations to the communities. From this perspective, individuals can be classified as (see Figure 1 for illustrations):

1) with no affiliation to any community;
2) connecting multiple communities;
3) important members of a community;
4) ordinary/majority members of a community;
5) non-important members of a community, who do not noticeably affect the community.

Based on this classification, we define the following four fundamental roles. The framework we present can be extended to include other specific roles based on a particular application, following similar methodology we present here. As the basis, we limited our focus to the most cross context roles.

LEADERS are the outstanding individuals in terms of centrality or importance in each community. Leaders are pioneers, authorities, or administrators of communities.

OUTERMOSTS are the small set of least significant individuals in each group, whose influence and effect on the community are below the influence of the majority of the community members.

MEDIATORS are individuals who play an important role in connecting communities in a network. They act as bridges between distinct communities.

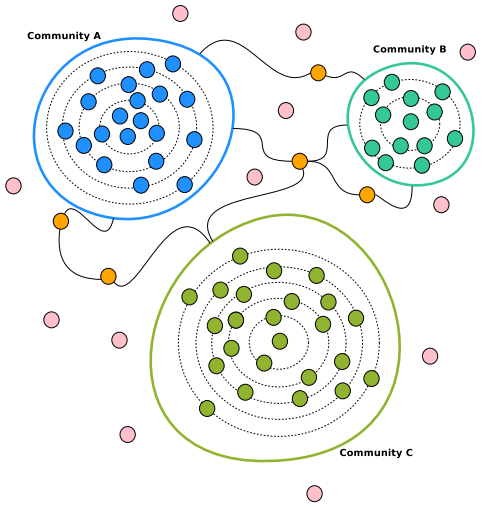


Fig. 1: An intuitive illustration for different types of roles based on structural properties, community affiliations, and members position within communities. In this figure, three communities A, B, and C are shown. Nodes are colour coded based on their role and affiliation: orange represents nodes that are connecting communities to each other (These nodes might also be part of community, however, we have ignored that case in this figure for simplicity, but considered in our definitions.); pink represents nodes with no connections or very week connections to communities; members of community A, B and C are coloured blue, dark green, light green respectively. Within each community, nodes are positioned based on their importance, i.e. closer to the borders of communities, the weaker and more inactive they are.

OUTSIDERS are individuals who are not affiliated to any one community. They either have almost equal affiliation to different communities, or have very weak ties to a community. The latter are commonly reffered to as outliers, whereas the former are exclusive mediators.

STRUCTURAL SOCIAL ROLE IDENTIFICATION

Having the definition of these four fundamental roles, we now describe how they can be identified in a given network. These roles and their identifications are based on a set of (non-overlapping) communities, which provide an intermediate representation for the structure of the network. The role mining process assumes that the communities are given, either explicitly known, or extracted by a community mining algorithm. Having the communities, it identifies roles either directly based on the community memberships (outsiders), or based on a ranking of nodes within the communities (leaders and outermosts), or the whole network (mediators). For the ranking based roles (i.e. leader and outermost), the distribution of the ranking scores for nodes is used to automatically identify the roles. Whereas, the identification of the mediator role, requires a more complicated procedure, described in the greedy *MedExtractor* Algorithm. More formally, we identify each of the four fundamental roles defined in the previous section as follows:

OUTSIDER members are the most straightforward role to identify. Having the communities of a network, individuals who do not belong to any community are identified as outsiders.

LEADER members are identified in association with each community. First, an appropriate importance/centrality measure *M* is used to score the members of the community (one might apply any of the commonly used centrality measures, or any other analysis that provides a ranking for importance of nodes). Then, the probability distribution function (pdf) for the importance scores (in that community) is estimated. Analyzing the characteristics of this pdf, determines the leaders. More specifically, nodes falling in the upper tail of the distribution, are identified as the community leaders.

OUTERMOST members are identified in contrast to the leaders, i.e. as members of a community falling in the lower tail of the importance pdf[1].

MEDIATOR members are the ones that connect different communities. Commonly-used betweenness centrality ranks nodes based on the number of shortest paths that pass through the nodes. Thus, a node with high betweenness centrality is a connectors between all nodes of the network without considering the notion of communities. Using this analogy, we propose two extensions of the betweenness centrality to include communities for extracting mediators in a network: *C-Betweenness* and *L-Betweenness*.

C-Betweenness counts the number of shortest paths between different communities that pass through a node. Let s_p and e_p denote the start point and end point of the shortest path p, respectively. Also let c_v return the community that node v belongs to. Now consider the set of all shortest paths that connect different communities as $CPaths = \{p \,|\, c_{s_p} \neq c_{e_p}\}$. Also let $I_p(p, v)$ return 1 if node v resides on path p, and 0 otherwise. Then, C-Betweenness of node v is defined as:

$$CBC(v) = \frac{1}{2} \sum_{p \in CPaths} I_p(p, v)$$

for undirected networks, where the division by 2 is omitted for directed graphs.

L-Betweenness considers only the shortest paths between leaders of different communities. Let $leaderSet(c)$ denote the set of leaders of community c, then consider $LPaths$ as:

$$LPaths = \{p \in CPaths \,|\, \exists\, c_i, c_j : s_p \in leaderSet(c_i) \wedge e_p \in leaderSet(c_j)\}$$

L-Betweenness of a node v, $LBC(v)$, is then defined as:

$$LBC(v) = \sum_{p \in LPath} I_p(p, v)$$

Figure 2 illustrates the difference between the proposed LBC and CBC scores. According to the LBC score, node A gets higher score, however considering CBC scores, node B is more important. Thus, based on the structure of the network, its communities, and more importantly the application of the mediators, one of these measures reflects the importance of the nodes better than the other.

[1]One should note that identifying outermosts is challenging using centrality measures. Because the intuition behind centrality measures is to identify higher values, not lower ones. This means, high centrality scores for nodes infer their importance, however, low centrality scores do not necessarily mean that they are not important. Thus, in general, centrality measures are efficient for identifying more central nodes, but not necessarily least central ones.

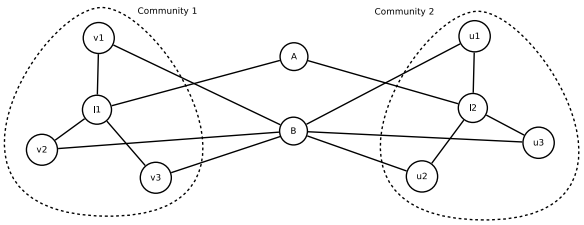


Fig. 2: This figure presents a synthetic network consisting of two communities. Leaders of two communities (l_1 and l_2) are connected to the node A, while other nodes are all connected to node B. Computing LBC and CBC for all nodes of the graph, the results are as follows: $LBC(A) = 1$, $LBC(B) = 0$, $CBC(A) = 7$, $CBC(B) = 12$, $\forall i : CBC(v_i) = CBC(u_i) = 6$, and $CBC(l_1) = CBC(l_2) = 3$.

The proposed CBC and LBC measures are necessary for identifying mediators, but are not sufficient. Consider a network consisting of 10 communities and two mediators M_1 and M_2 where both lie on 100 shortest paths between communities, and therefore have the same CBC values. Whereas, M_1 connects two distinct communities, while M_2 connects all 10. In this scenario, M_2 is connecting communities more globally and should be considered more of a mediator compared to $M1$. Therefore, in our role mining framework, we also incorporate the notion of **diversity score**, which indicates the distinct communities that are connected through a mediator. We define two variants for the diversity score: DS_{count} and DS_{pair}.

$DS_{count}(v)$ is simply defined as *the number* of distinct communities connected through a node. Let $I_d(c_i, v)$ to be 1 if $\exists\, p \in CPaths : s_p \in c_i \wedge v \in p$. Then DS_{count} of node v can be calculated as follows:

$$DS_{count}(v) = \frac{1}{2} \sum_{c_i} I_d(c_i, v) \quad \text{for undirected networks.}$$

(division by 2 is omitted for directed graphs.)

Diversity score can also be defined to count pairs of communities that have at least one shortest path between their members passing through node v. This variant of diversity score is denoted by $DS_{pair}(v)$. Prior to definition of DS_{pair}, we define $I_d(c_i, c_j, v)$ such that it is 1 if $\exists\, p \in CPaths : s_p \in c_i \wedge e_p \in c_j \wedge v \in p$. Then we have:

$$DS_{pair}(v) = \frac{1}{2} \sum_{c_i} \sum_{c_j \neq c_i} I_d(c_i, c_j, v) \quad \text{undirected graph.}$$

(division by 2 is omitted for directed graphs.)

In both DS_{count} and DS_{pair}, we can consider the shortest paths between nodes. Thus, a connection between two communities means having at least one shortest path between their members.

Using a combination of the above metrics, the mediacy score of nodes can be computed as a function of LBC or CBC and DS_{count} or DS_{pair}. We rank nodes by their mediacy score and propose *MedExtractor* algorithm to identify highly ranked nodes connecting the maximum number of communities as mediators.

Algorithm 1 MedExtractor: Find Mediators from SortedList based on their Mediacy Score

```
1: procedure ExtractMediators (Graph G, OrderedList L)
2:     ▷G is the graph associated with a network
3:     ▷L is descending OrderedList containing nodes of the network sorted based on their mediacy score.
4:     mediatorSet = {}          ▷ set of selected nodes as mediators
5:     connectedComs = {}        ▷ set of communities connected to eachother by nodes in mediatorSet
6:     while connectedComs.size < G.CommunityCount do
7:         n ← L.top()
8:         for all Community c ∈ n.incedentCommunities() do
9:             if c ∉ connectedComs then
10:                Add n to mediatorSet
11:                Add c to connectedComs
12:            end if
13:        end for
14:        L.remove(n)
15:    end while
16: end procedure
```

CASE STUDY

We apply the SSRM framework to identify roles in the *Enron email network* which contains emails exchanged between employees of the Enron Corporation. Without loss of generality, we study the year 2001, the year the company declared bankruptcy, and only consider people who had sent at least one email per day to filter out non-informative nodes. The resulting graph has almost 250 nodes and 1500 edges, with each month being one snapshot. Since the SSRM framework is built upon the assumption of the existence of communities, we apply the computationally effective local community mining algorithm [14] to produce sets of disjoint communities for each twelve snapshots. The Enron dataset is dynamic, hence, its extracted communities evolve and change over time. To capture the changes of the communities over the observation time, we apply the community events proposed by Takaffoli et al. [15] as the underlying temporal events. These events are later used in the analysis to observe the mutual effects between the changes in the extracted roles and the community events. We expect that the changes in the individuals role, play an important factor in the evolution of their corresponding communities.

Choosing a Centrality Measure

Intuitively, degree and closeness centrality scores seem to be good candidates for ranking individuals in a community in order to identify leaders and outermosts. However, as shown in Figure 3, degree distributions have mostly one tail (the upper tail). Although the long upper tail of degree distributions can be used to identify leaders, outermosts cannot be simply identified using degree distributions of communities. Thus, we consider closeness centrality rather than degree centrality that follows a normal-like distribution (Figure 4); the two tails in each community can be effectively used to identify both leaders and outermosts.

We use measures defined in the previous section to compute the mediacy score of a node. The number of leaders and the number of shortest paths between them are small in

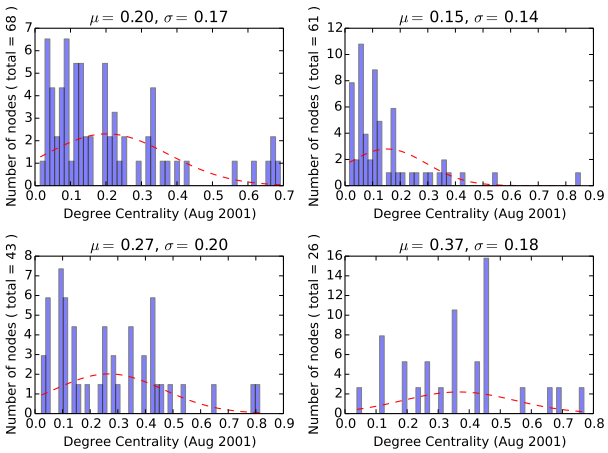


Fig. 3: Degree distribution of four community in August 2001.

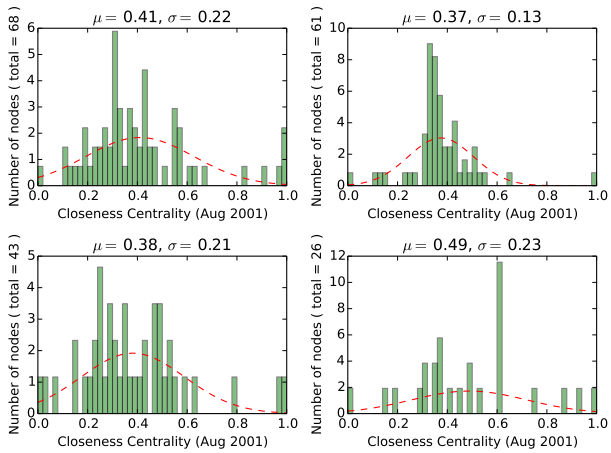


Fig. 4: Closeness distribution of four community in August 2001.

Enron dataset, thus, LBC is not an appropriate indicator for mediators. Hence, in this experiment, we use CBC to identify mediators. Although CBC is more expensive in terms of time complexity, it identifies mediators that could not be identified by LBC. Mediators found by CBC more strongly connect members of two communities, while mediators found by LBC connect community leaders. In addition to time complexity, CBC has another challenge; the probability of finding more prominent mediators between larger communities is higher in comparison to the smaller communities. This situation happens because there are more members in larger communities that results in having more shortest paths between them. To compensate for this effect, we use *normalized CBC* as follows:

$$NBC(v) = \frac{1}{2} \sum_{p \in CPaths} \frac{I_p(p, v)}{min(|c_{s_p}|, |c_{e_p}|)}$$

where division by 2 is omitted for directed graphs.

While two mediators may have equal scores, they can substantially be different in terms of the number of distinct communities they connect. Thus, we also consider diversity score which takes into account the variant of communities. Based on this idea, we define *mediator score* as the multiplication of the node's normalized C-Betweenness and diversity score as follows:

$$MS(v) = NCB(v) \times DS_{count}(v). \quad (1)$$

Identifying Roles

After choosing a metric to compute nodes' score, we develop the following methods to extract roles from the ranked lists.

The closeness distribution used for identifying leaders and outermosts in the Enron dataset is close to the normal distribution. Using properties of the normal distribution, we set $\mu + 2\sigma$ as the upper and $\mu - 2\sigma$ as the lower thresholds to identify leaders and outermosts respectively. Figure 5a shows communities of Enron in the August 2001 and visually presents how central a node is within its community. Other than this, more information from distributions such as gaps could also be used in identifying these roles.

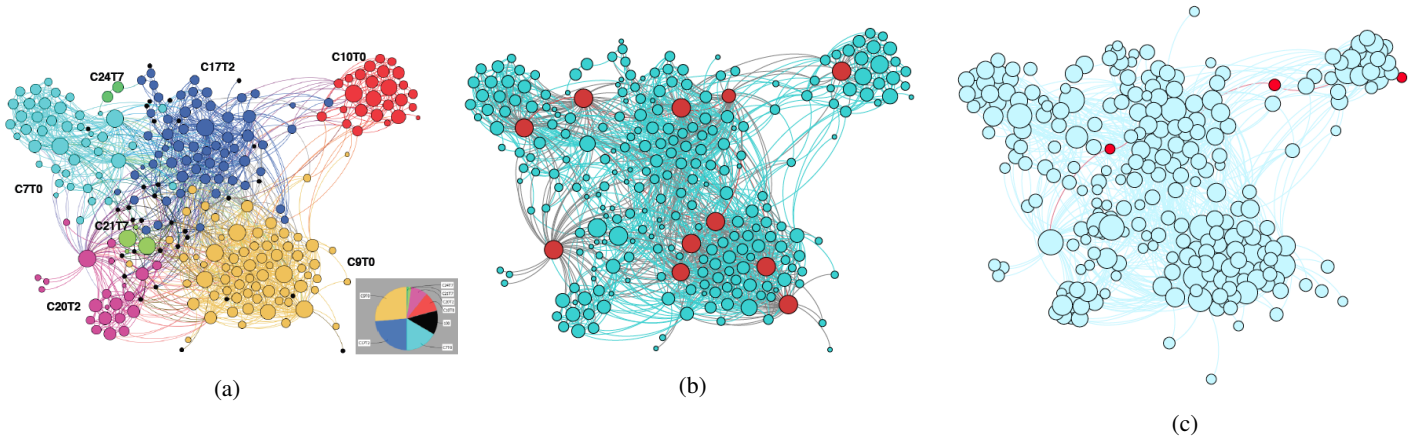
Email	Community	Position
jeff.dasovich@enron.com	C9T0	Executive/Director for State Government Affairs
giger.dernehl@enron.com	C9T0	
richard.shapiro@enron.com	C9T0	VP regulatory affairs (Enron's top lobbyist)
kimberly.watson@enron.com	C10T0	Director
d..steffes@enron.com	C9T0	Vice President
kenneth.lay@enron.com	C17T2	CEO, chairman, and chief executive officer
e..haedicke@enron.com	C7T0	Managing director
susan.mara@enron.com	C9T0	California director of Regulatory Affairs
billy.lemmons@enron.com	C17T2	Vice President
becky.spencer@enron.com	C7T0	
l..denton@enron.com	C20T2	Lawyer

TABLE I: Leaders of the network shown in Figure 5b, their community affiliation and position in the Enron organization

Leaders identified in August 2001 are shown in Table I. Being a leader in the network (Figure 5b) translates into having a high average of short email distance to other individuals in the community. Among these names, Kenneth Lay is founder, CEO, chairman, and chief executive officer of Enron. Kimberly Watson is likely to be one of the influential directors in the company based on the information we found in two meeting minutes [16], [17].

Having the ranked list of nodes based on their mediator score, we need a method to extract mediators. To this end, we use and compare two methods. In the first one, we use the MedExtractor (Algorithm 1), a greedy algorithm to identify mediators. MedExtractor tries to minimize the number of mediators that connect the maximum number of communities in a network. Mediators found by this algorithm are high score individuals according to their mediator score that all together maximize the number of connected communities in the network. These nodes identified as mediators are shown in red in Figure 6b.

In the second method, we use the properties of the distribution of mediator score to identify mediators. The mediator score distribution is close to a power-law distribution. This means a small number of nodes (the ones in the tail of the distribution) are in control of lots of shortest path between communities in our dataset. Based on this property, we use the tail of the mediator score distribution to identify mediators. However, the tail of the mediator score distributions for the Enron dataset is very sparse. The reason for sparsity of the



(a) (b) (c)

Fig. 5: (a) communities within the Enron email network in August 2001. Colours represent communities except for black that represents outsiders. The bigger the size of the node, the more central it is in the community. (b) Leaders of communities are shown in red, (c) outermosts are shown in red. Only two communities have nodes serving as outermosts. From the right to left, outermosts are Ava Garcia (probably an assistant according to the body of some emails), Shirley Crenshaw (probably an assistant), and Leslie Reeves a module manager.

Email	Community	MedExtractor	Mediator Score
kenneth.lay@enron.com	C17T2	✓	1
kam.keiser@enron.com	C17T2	✓	0.486
l..denton@enron.com	C20T2		0.378
gerald.nemec@enron.com	C7T0		0.338
janette.elbertson@enron.com	C7T0		0.332
kimberly.bates@enron.com	C7T0	✓	0.294
e..haedicke@enron.com	C7T0		0.270
tim.belden@enron.com	C20T2		0.193
jeff.dasovich@enron.com	C9T0		0.181
rika.imai@enron.com	C21T7		0.180
d..steffes@enron.com	C9T0		0.179
w..cantrell@enron.com	C9T0		0.175

TABLE II: Top-12 nodes (based on their mediator score) as mediators using gaps in the distribution of mediator score.

	Jan-01	Feb-01	Mar-01	Apr-01	May-01	Jun-01	Jul-01	Aug-01	Sep-01	Oct-01	Nov-01	Dec-01	
jeff.dasovich@enron.com	0	0.91	0.97	1	0.91	0.86	1	1	0.94	0	0	0	8
tana.jones@enron.com	1	1	1	1	1	0	0	0	0.89	1	0	0	7
james.steffes@enron.com	0.82	1	0.85	0.73	1	0.79	0	0			0		6
richard.shapiro@enron.com	0						0.82	1	0.99	1	0.86	0.87	6
d..steffes@enron.com	0							0.98	0.94	0.96	1	0.85	5
marie.heard@enron.com	0						1	0	1	0.94	1	0.78	5
becky.spencer@enron.com	0			0.98	0	1	1	0.96	0	0	0	0	4
louise.kitchen@enron.com	0								1	0.73	1	1	4
ginger.dernehl@enron.com	1	0	0	0	0.86	0	0	1	0	0	0	0.87	4
susan.mara@enron.com	0					0.9	0.79	0.91	0.94	0	0	0	4
kathryn.sheppard@enron.com	0								1	1	1	1	4
alan.comnes@enron.com	0		0.84	0	0	1	0	0	0.92	0	0	0	3
mary.hain@enron.com	0.84	0.92	1	0	0								3
janel.guerrero@enron.com	0			0.73	0	0	0	0	0	1	0	1	3
john.lavorato@enron.com	0			1	1	0	0	0	0	0	0.91	0	3

Fig. 7: Changes of nodes serving as leaders in the Enron dataset.

tail could be due to the small population of the network. To overcome this problem in our dataset, we use the point where the tail of the distribution starts getting sparse (the first or second gap in the histogram) as the lower threshold to identify mediators. Table II presents the top-12 mediators chosen considering the sparseness and gaps in the distribution. The nodes that are identified as mediators by this method, are not only connecting communities, but also reside on most of the shortest paths between communities.

MedExtractor determines the minimum number of mediators that connect the maximum number of communities in the network, while using information from distributions, we find nodes controlling most of the connections between communities. As shown in Figure 6b and Table II, nodes that are selected by the MedExtractor are a subset of nodes identified by the analysis on the distribution function. Thus, depending on what we expect from mediators, one of the aforementioned methods can be used.

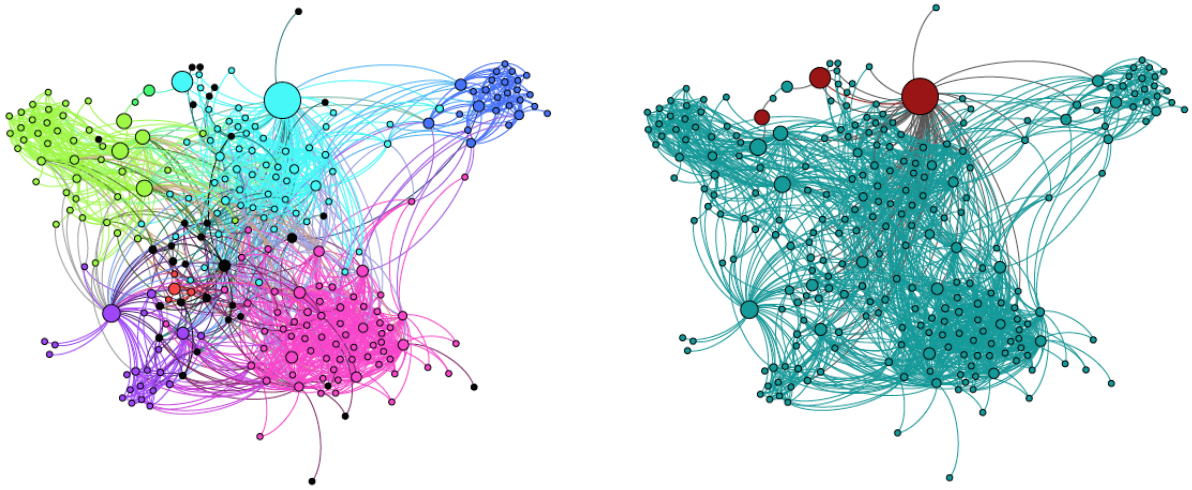
Roles Changes

In this section, we present the results on how nodes change their role through time. Since two important roles in our proposed framework are leaders and mediators, the focus in this section is on nodes that have been leaders or mediators at least once in different timeframes.

We present change of roles by means tables in Figure 7 and Figure 8, where rows and columns are respectively nodes and timeframes. The intersection of rows and columns has information about the role of the associated node in the respective timeframe. For better visualization, cells are coloured to represent nodes' relative strength in having a specific role.

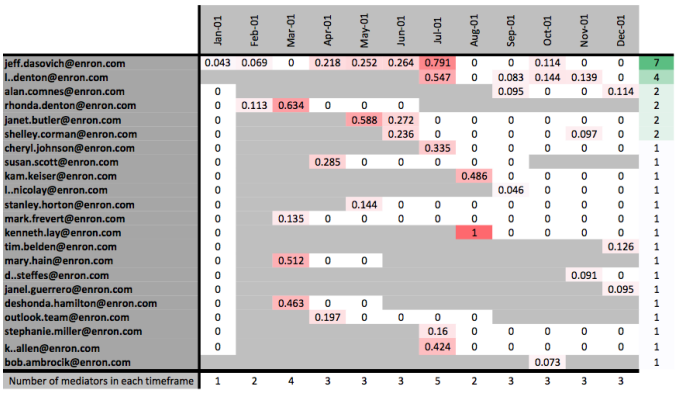
We track nodes' roles after the first time they become leaders. Thus, whether they are present in the network or not before their first become a leader is not depicted in Figure 7. This figure provides us with interesting information about leader, i.e. some nodes are constantly leaders in early timeframes while some others are constantly leaders in later timeframes. Also, the importance of the leadership of leader nodes change over time. Intuitively, being a constant leader over more timeframes could mean that a node is more important and influential in its community.

Figure 8 presents how mediators change their role through time. Unlike the leading role where several nodes attain their leadership over several timeframes, we observe more fluctuations over the mediating role over time. According to Figure 8, mediators identified for the months July, August,



(a) Communities (b) Mediators

Fig. 6: a) Communities within August 2001 timeframe. Size of the nodes depicts their mediator score. b) mediators (red nodes) found by MedExtractor



	Jan-01	Feb-01	Mar-01	Apr-01	May-01	Jun-01	Jul-01	Aug-01	Sep-01	Oct-01	Nov-01	Dec-01	
jeff.dasovich@enron.com	0.043	0.069	0	0.218	0.252	0.264	0.791	0	0	0.114	0	0	7
l..denton@enron.com							0.547	0	0.083	0.144	0.139	0	4
alan.comnes@enron.com	0								0.095	0	0	0.114	2
rhonda.denton@enron.com	0	0.113	0.634	0	0	0							2
janet.butler@enron.com	0				0.588	0.272	0	0	0	0	0	0	2
shelley.corman@enron.com	0					0.236	0	0	0	0	0.097	0	2
cheryl.johnson@enron.com	0						0.335	0	0	0	0	0	1
susan.scott@enron.com	0			0.285	0	0	0	0	0				1
kam.keiser@enron.com	0							0.486	0	0	0	0	1
l..nicolay@enron.com	0								0.046	0	0	0	1
stanley.horton@enron.com	0				0.144	0	0	0	0	0	0	0	1
mark.frevert@enron.com	0		0.135	0	0	0	0	0	0	0	0	0	1
kenneth.lay@enron.com	0							1	0	0	0	0	1
tim.belden@enron.com	0											0.126	1
mary.hain@enron.com	0		0.512	0	0								1
d..steffes@enron.com	0										0.091	0	1
janel.guerrero@enron.com	0											0.095	1
deshonda.hamilton@enron.com	0		0.463	0	0								1
outlook.team@enron.com	0			0.197	0	0	0	0					1
stephanie.miller@enron.com	0						0.16	0	0	0	0	0	1
k..allen@enron.com	0						0.424	0	0	0	0	0	1
bob.ambrocik@enron.com									0.073				1
Number of mediators in each timeframe	1	2	4	3	3	3	5	2	3	3	3	3	

Fig. 8: Changes of nodes serving as mediators in the Enron dataset.

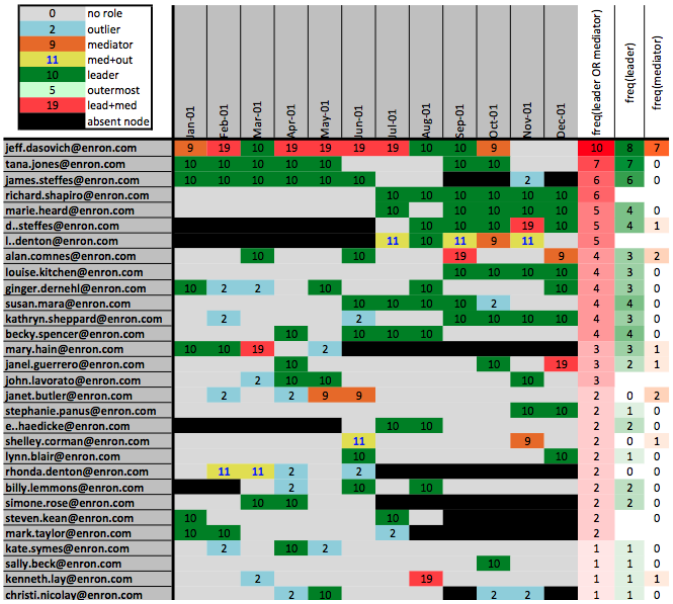


Fig. 9: Role changes in the Enron email network.

and September express higher mediator scores. This shows that more conversations have happened between communities of the Enron network in these three months.

Finally, Figure 9 depicts how nodes change their roles through time. Note that this figure presents the information merely relevant to the nodes that have served as leader or mediator in at least one timeframe. Based on the information presented in this figure, there are examples that a node simultaneously acts as both a mediator and a leader. This figure provides knowledge regarding the role transition of each node. For example, Jeff Dasovich is a mediator in January 2001, then transforms into simultaneous leader and mediator. Later, he becomes only a leader and then regains his leading and mediacy until October. He is not a leader any more in October and does not have any role after that time. One other example is Kenneth Lay who is an outermost in March and later on becomes both a leader and a mediator in August. This role change is coincident with the time when the former Enron CEO, Jeffery Skilling, resigned and Lay became the new CEO. Tracking such temporal changes results in interesting information that helps in analyzing phenomena happening in the network. We use these transitions to analyze events happening for communities.

Role Transitions and Community Events

Tracking role changes in aggregation with community events in the Enron email network provides interesting result on the impact of role changes on the community changes.

Steven Kean, Enron executive vice president and chief of staff, is a leader in community C9T0 in August but no more in

Dataset	t_{BC}	t_{CBC}	t_{LBC}	t_{BC}/t_{CBC}
karate club (34 nodes)	64_{ms}	118_{ms}	57_{ms}	0.54
Enron Oct. 2001 (228 nodes)	1012_{ms}	762_{ms}	152_{ms}	1.33

TABLE III: Running time comparison between BC, CBC, and LBC on the karate-club and Enron networks.

September. Coincidentally, community C9T0 *dissolves* in the end of September.

Jeff Dasovich is both a mediator and a leader in C9T0 in March, but is not a mediator any more in April. Interestingly, community C9T0 faces a *merge* in the beginning of May with another large community.

Some other interesting role changes that happen with the event *split* are: Jan Moore who becomes a leader in C10T0 in March resulting by a *split* happening in C10T0 in April. Same thing happens for Becky Spencer who becomes a leader of C7T0 in April, and a month later a *split* occurs in C7T0. Similar scenarios happen for Alan Comnes in C9T0, Susan Mara in C9T0, and Ginger Dernehl in C9T0.

BC, CBC, and LBC Comparison

In this section, we present the comparisons between Betweenness centrality (BC), C-Betweenness centrality (CBC), and L-Betweenness centrality (LBC). We have computed these three metrics on two different networks: (i) karate-club network, and (ii) Enron communication network. We compare the top-k lists for these metrics. According to what we have found, more than half of the nodes are shared between the top-20 nodes for BC and CBC. Moreover, LBC for the karate-club identifies exactly those nodes that are important in connecting two communities. We also compare the running time to compute these metrics (Table III) on each of our networks. The algorithm used for computing BC scores is the one implemented in the JUNG framework and thus much more efficient than the ones implemented by us for CBC and LBC. The results on the running time shows however, BC was implement more efficiently and performed much better than CBC on the karate-club, CBC is better than BC when size of the network increases. This shows the growth of BC is more than CBC and LBC.

Conclusion and Future Work

We developed a framework for structural role mining in social network called SSRM in this paper. In order to evaluate the SSRM framework, we applied it on the Enron communication network. We identified outsiders, outermosts, mediators, and leaders. Among these roles, leaders and mediators are the important ones to study more. Thus, we found information about the people associated with nodes having the role of a leader or a mediator. According to our findings, they are indeed important people in the Enron organizational hierarchy. Moreover, we observed how nodes change their role through time. Based on role changes, we analyzed community events happening in the Enron network and observed mutual relation between role changes and community events.

Tracking how these roles change through time provides information about the temporal characteristics of nodes and the network. Hence, defining dynamic roles by focusing on temporal patterns of role changes can be considered as a future work to extend the SSRM framework.

References

[1] B. J. Biddle, "Recent development in role theory," *Annual Review of Sociology*, pp. 67–92, 1986.

[2] J. Walton, "Differential patterns of community power structure: An explanation based on interdependence," *The Sociological Quarterly*, vol. 9, pp. 3–18, 1968.

[3] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1979.

[4] O. A. Oeser and F. Harary, "A mathematical model for structural role theory: I," *Human Relations*, 1962.

[5] O. Oeser and F. Harary, "A mathematical model for structural role theory: II," *Human Relations*, 1964.

[6] O. Oeser and G. O'BRIEN, "A mathematical model for structural role theory: III," *Human Relations*, 1967.

[7] M. Forestier, A. Stavrianou, J. Velcin, and D. A. Zighed, "Roles in social networks: Methodologies and research issues," *Web Intelligence and Agent Systems*, vol. 10, no. 1, pp. 117–133, 2012.

[8] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: structure and algorithms," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 221–230.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Technical Report 1999-66, November 1999.

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

[11] E. S. Kim and S. S. Han, "An analytical way to find influencers on social networks and validate their effects in disseminating social games," in *International Conference on Advances in Social Network Analysis and Mining (ASONAM)*. IEEE, 2009, pp. 41–46.

[12] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM)*. ACM, 2008, pp. 207–218.

[13] J. Scripps, P.-N. Tan, and A.-H. Esfahanian, "Node roles and community structure in networks," in *Proceedings of the 9th WebKDD and 1st SNA-KDD workshop on Web mining and social network analysis*. ACM, 2007, pp. 26–35.

[14] J. Chen, O. Zaïane, and R. Goebel, "Local community identification in social networks," in *International Conference on Advances in Social Network Analysis and Mining (ASONAM)*. IEEE, 2009, pp. 237–242.

[15] M. Takaffoli, J. Fagnan, F. Sangi, and O. R. Zaïane, "Tracking changes in dynamic information networks," in *International Conference on Computational Aspects of Social Networks (CASoN)*. IEEE, 2011, pp. 94–101.

[16] "ENRON:Minutes for August," http://datasets.opentestset.com/datasets/Enron_files/full/watson-k/MINUTES_FOR_AUGUST.doc.

[17] "ENRON:Minutes for July," http://datasets.opentestset.com/datasets/Enron_files/full/watson-k/MINUTES_FOR_JULY.doc.