University of Alberta

STRUCTURAL ROLE MINING IN SOCIAL NETWORKS

by

Afra Abnar

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

Spring 2014
Edmonton, Alberta

Abstract

All along their lives, individuals take roles in their interactions with each other. This behaviour is known as the *role-taking* characteristic of human beings. We refer to these roles as *social roles* that are the primary components of societies. Identifying social roles in a society helps to better analyze the social phenomena. Consequently, it can be beneficial in the search for influentials, trustworthy people, idea innovators, etc. With this intention, we propose the *structural social role mining (SSRM)* framework to identify roles, study their changes, and analyze their impacts on the underlying social network. More specifically, we define four fundamental roles called *leader*, *outermost*, *mediator*, and *outsider*. Subsequently, we suggest methodologies to identify these roles within a social network. While exploring our proposed methodologies for identifying the aforementioned roles, we develop two new variants of Betweenness centrality (BC): *LBetweenness (LBC)* and *CBetweenness (CBC)*. Motivated by time complexity, these two centrality measures are computed more efficiently compared to *Betweenness centrality* especially in large social networks. Eventually, we identify and study changes of roles in the *Enron communication network* using our proposed framework. According to our results, individuals serving as leaders or mediators were important people in the Enron organization. Moreover, identifying roles as well as their changes through consecutive timeframes could be informative and thus could be used as a clue for further investigations.

Acknowledgements

We do not always recognize, but there are wonderful people supporting us in traveling each stage of our life. This thesis is done with all guidance and supports of the amazing people around me.

First and foremost I offer my sincerest gratitude to my supervisor, Dr. Osmar Zaïane, who has supported me throughout my thesis with his patience, ideas, and knowledge whilst allowing me the room to work in my own way. I really appreciate his help leading me in finding my way through out the research. I attribute the level of my Masters degree to his encouragement.

In my daily work I have been blessed with two friendly and cheerful fellow students. Reihaneh Rabbany, who helped me not only in the first steps of my research but also throughout my work with useful discussions, always sent me useful and interesting links and was always there whenever I faced problems or needed help. She is a wonderful friend and offered her help to me at any time. Mansoureh Takaffoli, whom I unfortunately had the chance to work with for the last month of my research but taught me how should I deal with different challenges throughout the research and push it to get progress.

The story of coming to UofA happened to me when one of my best friends Neda Mirian got an offer and moved here. The next year my other best friend Parisa Delfani came to UofA and that directed me to apply for post graduate studies at UofA. Thanks to my dear friends, Neda and Parisa for being here and helping me not feeling alone with their supports. Living in the cold city of Edmonton resulted in gradually finding many warm and awesome friends. Special thanks to Moslem Noori, my kind friend, for his supports especially in the last hard weeks of writing my thesis with reading my thesis and his thoughtful comments. People that I am

so lucky to have the chance to be friend with are Saeed Mohajeri, that I would like to thank him for his great company not only as supportive friend, but also when thinking on a problem related to my research or trying manage time when I was in rush, Mohsen Taghaddosi for all his supports and our great discussions, and Farzaneh Mirzazadeh for being a really good friend. Good friends are gifts of life that make life sweater and happier and I am so lucky to have a bunch of them around.

Last but not the least, I could have never kept moving forward without supports of my beloved parents, Noushin Sharif and Aziz Abnar. I would like to thank them millions of times for their efforts and helps in every stage of my life. Also my dear sister, Samira, and awesome brother, Ali, who always positively encourage me in my life and specially in my work. I owe whatever I have and gain to my wonderful family.

There are many many other people in Edmonton, Tehran, and many other cities in this world who should be named here. I appreciate how my friends always support me and I am so pleased to know and be friend with these amazing people.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

When you don't have many friends and you don't have a social life you're kind of left looking at things, not doing things. There's a weird freedom in not having people treat you like you're part of society or where you have to fulfill social relationships.

Tim Burton

Human beings have lived in forms of societies and communities throughout their existence to overcome difficulties and barriers by working collaboratively. In the early days, these societies were in the form of families and small tribes. Later on, societies evolved to more complex forms during different stages of civilizations and concepts such as villages, cities and countries appeared to describe new forms of human societies.

Regardless of the complexity of societies, they essentially consist of *members* and *relations* between these members. These relations determine the structure of each society and how members of a society collaboratively work with each other to achieve a common goal or gain mutual benefits. As a result, studying the patterns of these relations and the structure of a society can lead us to understand how a society may succeed or evolve as the time passes on. Traditionally, *sociologists* have focused on studying the structures of human societies. They have investigated interactions among individuals and how these structures affect phenomena happening in the society as well as the behaviour of individuals. However, with the advent of new technologies and internet, *social network analysis* has become a fundamental *interdisciplinary* science.

The exploding number of people who are involved in online social networks has brought many aspects of human life into the cyber world. Nowadays, everyone has at least heard of one of social networking websites or even has gone further and actively uses them. The original purpose of the first social networking websites was to help people to stay connected with their friends. However, as the popularity of these websites skyrocketed, many applications started to launch on social networking platforms. As a result, the online social networking contains a vast amount of information not only about the friendship and relation between people, but also other aspects of their lives including habits, dreams, achievements, etc. This has even made social networking websites a better representative of the relations between people in the real world. Thus, more information about our society than any other time is now in forms of 0s and 1s which makes computer scientists involved in the study of societies beside socialists and psychologists.

1.1 Motivation

There exist many challenging questions when it comes to studying human behaviours. Researchers are engaged in problems like finding groups of friendship or clusters of closer individuals, identifying idea innovators, detecting powerful individuals who affect others' beliefs and behaviours.

Individuals are playing a significant role in the phenomena happening in society, as they are basic components of social networks. Historically, it was an elder in a tribe who had the power to affect faith of his people. Nowadays we have political leaders, influential friends, trustworthy colleagues, and innovators in our society who have tremendous impact on others' lives and the whole society. That is why many researchers have focused on understanding characteristics of individuals in terms of how they are connected to each other and how structurally they are positioned in the network as well as their behaviours and actions. For instance, in a friendship network between high school students, how a student can encourage her classmates to buy a special brand of stationery, what are her actions, and how is she connected to the rest of the students are important factors in making a student

influential among his classmates.

To formally define the influence of network members the concept of *role* has been introduced in sociology. The concept role refers to a special position, which is associated with a set of behaviours, expectations, and responsibilities. People can have roles in their communities and societies. Thus they are having special behaviours and others expect certain actions from them. Having roles empower people in affecting their societies. From an analyst's point of view, role is a concept that helps us better understand individuals' behaviours and their interactions in the society.

Thanks to online social networking tools, research on human networks differs from traditional ones. With the large amount of data available nowadays, we need new techniques to explore human relations. From a social network analysis perspective, various metrics are defined for structural and non-structural (behavioural) characteristics of individuals which enable us to find roles in this huge amount of data. Using these metrics, we come up with the idea of "Social Roles" to get a more high-level insight on individuals' characteristics. Roles can be defined based on values of one or more metrics to map data-driven and graph-theoretic concepts from social networks to real-life ones.

Based on metrics defined to measure characteristics of networks and individuals in social networks, we introduce "Social Role Mining", a framework to unify these metrics into the high-level concept of role. Roles can give more insightful meanings to the raw metrics, if defined well on social networks. Defining and identifying roles can also help researchers to better analyze events taking place in a network. Moreover, roles can be used to simplify complex networks. Thus, they may even lead to more efficient algorithms for large social network graphs.

1.2 Challenges

To study roles and their effects in a social network, we face several substantial challenges listed as follows:

- *Role definition:* there is still no consensus on the definition of role [10] in

sociology that we can directly borrow in our work. For example, the role teacher in a classroom, or the role influential among students of that classroom are two different roles people take in society. A prominent question in this thesis is: *what do we mean by the word "role" [13] and how do we transfer definitions from social science to computer science.* Having a reasonable definition for the concept of role, we should answer the following questions: (i) how do we define different types of roles in a social network, (ii) what kinds of roles are of our interest to define, (iii) what kinds of properties from the network do we take into account to define roles.

- *Roles and temporal changes:* the other important issue is time. Since time is an intrinsic property of social networks, how roles should be defined to reflect the effects of time. In other words, how roles may change if the effect of time is ignored in their definition.

- *Benefits of roles:* finally, having a set of meaningful roles defined on a social network, how can they be used to simplify or model the underlying social network. On the other hand, how algorithms can use these roles to perform more efficiently on large social networks.

1.3 Thesis Statement

In this thesis we focus on the following statements:

St 1. The concept of *role* from social science [10] can be defined as an analysis means in network science.

St 2. The concept of *role* can unify various methodologies focusing on studying individuals' positions and importance in a network.

St 3. Measures and metrics defined in network science can be used to identify more insightful concepts in a network.

St 4. There are ways to more efficiently measure characteristics of nodes in large social networks by deeper analysis on the network.

1.4 Contributions

In this thesis, we contribute to social network analysis by proposing a framework for studying individuals' behaviours and their relations. The proposed framework is built using the concept of *role* from social science. We define the concept of "role" in terms of social network terminology. Moreover, the framework unifies various kinds of metrics like centrality measures, and analysis such as influence, spread of innovation, trust, leadership, etc.

Our proposed framework in this thesis is called "Structural Social Role Mining" (SSRM). This framework defines a set of fundamental roles and proposes methodologies to identify those roles. Moreover, it is a context-independent framework that can be applied to any kind of social network as long as it is modeled as a homogeneous graph, where there is only one kind of nodes and one kind of edges (can be either directed or undirected, weighted or unweighted).

The term *structural* refers to the fact that only structural properties of the network, modeled as a graph, are considered in our analysis. This nomination is based on the perspective of categorizing the information in a network into structural and non-structural (behavioural) forms. Structural information relates to the structure of the network and the former refers to any other information which is associated with the graph representing the network and its elements.

The proposed framework is built upon two assumptions about the social networks. These two assumptions are basically the characteristics of human societies. The first assumption is that social networks are composed of multiple sets of highly connected nodes (communities). The second one, is the role-taking behaviour of individuals within a social network. Under these assumptions, we define the set of four fundamental structural roles: (i) leader, (ii) outermost, (iii) mediator, and (iv) outsider. These roles are defined based on structural properties of a network as well as the information about communities in a social network graph.

Subsequently, we provide methods to identify the aforementioned roles. Having the roles identified in a network, we study changes of roles in consecutive timeframes. With this in mind, we define *role events* capturing changes that happen in

roles of individuals through time. Moreover, we study the relation between these events and other structural events in a social network. These studies give us insight about the evolution of the network. We could also observe how impactful are individuals and their roles in relation to the changes that happen throughout the network.

Through defining the SSRM framework, we present efficient ways for computing metrics on large social graphs. We define two new varients of Betweenness centrality called *LBetweenness (LBC)* and *CBetweenness (CBC)*. Due to consideration of only a subset of shortest paths in a network, these two centrality measures are computed more efficiently compared to *Betweenness centrality (BC)*.

1.5 Limitations

Sociological phenomena usually cannot easily be modeled by the means of mathematics and therefore hard to study. In this work we study roles with the means of social network analysis. We have proposed the SSRM framework based on sociological facts. Thus, the work has its origin in sociology.

An important concept considered in this work, is the notion of *communities* in social networks. Similar to societies that are compounded of groups, social networks are also build up of communities. Working with data, most of the time communities are not given. Thus, we have to trust the results of community mining which is itself an open area of research in social network analysis. On the other hand, there is no ground truth for roles. Hence, the other limitation of this work is how to validate roles that are found by the SSRM framework. The validations are done based on other kinds of analysis and information we get out of the data. In fact, it is very hard to come up with ground truths for roles of individuals in a network. However, different attributes associated to nodes of a network can be used as means of validating the results.

With all these limitations, this work tries to reasonably study and model roles in the context of social networks analysis.

1.6 Thesis Organizations

The rest of this thesis is organized in five chapters. A brief overview of these chapters are as follows:

Chapter 2 presents the background work done in social network analysis and specifically in role analysis and role mining. In Chapter 2, we review works on social network analysis in general, their properties, characteristics and models. We see how social network graphs are different from other graphs. Then, we go further into metrics defined and used to measure various characteristics of social networks. We also present works on temporal social networks, community detection, and related research on social media. Later, we take a deeper look at what has been done in role analysis and role mining, the sociological point of view of roles along with works in data analysis and computer science.

Chapter 3 starts with the proposed framework for structural social role mining that defines roles and presents the intuitions behind the definitions. Since social networks are intrinsically dynamic, the framework suggests a set of fundamental roles in a single snapshot. The fundamental roles are: leader, outermost, mediator, and outsider. Outsiders are those with very weak connections to the rest of the network. Outermosts are the least important in the structure of a community and leaders are prominent individuals within each community. Finally, mediators are nodes that connect communities within a network. Having these fundamental roles identified, their changes in consecutive timeframes are studied as they might be significant in influencing the whole structure of the network or a part of it. Thus, studying the changes of roles through time leads to the identification of dynamic (temporal) roles.

To evaluate our proposed framework in Chapter 3, we present a case study on the *Enron communication network* Chapter 4. In studying the Enron communication network, we apply the SSRM framework on the dataset to identify the set of defined roles. Furthermore, since the dataset has timeframes, we are able to see the changes that happen in roles. Afterwards, we analyze the impact of changes in these roles

on the network.

Chapter 5 presents our investigations on the two variant of Betweenness Centrality (BC) called LBetweenness (LBC) and CBetweennes (CBC) that are defined in Chapter 3. In this chapter we compare the results of BC, LBC, and CBC on the karate club and Enron networks.

In Chapter 6, we conclude the work by reviewing our findings and discussing the results. We also present the applications of role mining and in particular our proposed framework along with possible future works.

Chapter 2

Background and Related Work

I read somewhere that everybody on this planet is separated by only six other people. Six degrees of separation between us and everyone else on this planet. The President of the United States, a gondolier in Venice, just fill in the names. I find it extremely comforting that we're so close. I also find it like Chinese water torture, that we're so close because you have to find the right six people to make the right connection. I am bound, you are bound, to everyone on this planet by a trail of six people.

Six Degrees of Separation by John Guare

2.1 Social Network Analysis

Networks of humans are complicated and an important subject to study. People are connected to their friends, colleagues, acquaintances, or connected by other myriad kinds of relationships. They form social groups, organizations and other structures in which they interact with each other, affecting the whole society and causing many events to happen. Thus, exploring the human beings networks and structures they build is a fundamental step in analyzing behaviours of the society as a whole and groups of people at different levels of granularity and even individuals themselves. Pool and Kochen in [18] have listed a number of questions and research directions in humans' networks.

A network in which entities are individuals and connections between entities represent individuals' relations is called a *social network*. Hence, a social network is made up of individuals (or organizations) and pairwise relations between them

as discussed. The study of social networks has been started since the early 1900s. Since social network analysis is about phenomena taking place in human society, groups, and organizations, it was originally a topic of study in sociology and psychology. In the 1930s, researchers in sociology started thinking about representing the shape of social structures. In this regard, in 'formal sociology' of Simmel, they use "points", "lines" and "connections" in order to analyze and describe social patterns. Impressed by the work in formal sociology, some psychologists have noticed the significance of the structure of a group in actions and behaviours of its members. Moreno in 1934, invented the terms 'sociogram' and 'sociometry' as a way of visually representing social networks with points and lines [61]. This is how the science of network gradually formed in its early days.

More precisely, social network analysis is an interdisciplinary field with its roots in sociology, psychology, statistics and graph theory. When it comes to applications of social network analysis, it encompasses broader range of topics including economics, politics, health sciences, marketing, etc. Moreover, from a general perspective, social networks are integrated within "Network Science". Network science is itself an interdisciplinary area of study focusing on complex networks such as computer networks, biological networks, neural networks, telecommunication networks and social networks. To put it in other words, the field of network sciences is fundamentally built up on social structure from sociology, graph theory from mathematics, data mining and information visualization from computer science and statistical mechanics from physics [78].

Social networks are represented by graphs where individuals of the network are represented by nodes, and similarly their connections are shown by edges. Therefore, various metrics are defined based on properties of nodes and edges in a graph. In this thesis, we use the words social network, social graph, complex network, network, and graph interchangeably as each social network has its corresponding graph.

2.2 Social Networks Properties

Talking to a stranger while you are waiting in a line or traveling on a plane, you may notice that you both know a specific person by first name. This scenario or a similar one, has happened to many people at least once. We may use the expression "What a small world!" in such a situation. Hence, there is a sense in society that the world is so small and people are connected through a small chain of their acquaintances. Pool and Kochen in 1960s, tried to mathematically study this phenomena. In their mathematical model, they studied the probability of two random people knowing each other and also the probability of them to be linked through their acquaintances. They had some obstacles in their studies to find reasonable answers to some of their questions. Altogether, some years later they published their model along with their unsolved questions in [18].

About the same time as Pool and Kochen, Stanley Milgram, an American social psychologist, designed experiments to study the average distance between American people [49, 73]. Later on, his experiments lead to the theory of *six degree of separation*. In one of the experiments they selected 296 random individuals from Nebraska and Boston as senders and a target. The senders were asked to forward a letter to the target if they know him in person or to an acquaintance they know by first name who they think might be closer to the target person. Although many letters did not reach the destination, among those who reached, the average number of intermediate individuals was between 5 and 6 approving the fact that there is a small distance between people in the society. This phenomena is called *small-world* characteristic.

More sophisticated analysis of the small-worldliness of our society was done by theoretic researchers by the means of graph theory. From a mathematical perspective, graphs could be ordered or random. But a vast majority of real world networks correspond with graphs that lie between these two extremes of being completely random and completely ordered. Watts in [75] investigates the small-world phenomenon as a general feature for *sparse*, *decentralized* networks which are neither completely random, nor completely ordered graphs. The world "small" in networks

means that every node is just a few hops away from any other node in the network. According to [75], the four characteristics that make small-world a surprising and remarkably important property of these network are as follows:

- Size of the network (number of nodes in the corresponding graph) should be numerically large. If size of the network is small, then the small distance between entities is an obvious fact, not an interesting one.

- The network should be sparse which means every node is directly connected to only a very small number of nodes. Consider entities have connections to almost all other entities in a network, then the small distance between them is again a straightforward conclusion.

- The network should be decentralized. It means that there should be no central node to which most of the other nodes are connected. This property concludes that the maximum degree of nodes in the corresponding graph is much less than the number of nodes. This condition is even stronger than the previous one which makes small-world more unbelievable.

- Finally, the network should be highly clustered. It means neighbours of a node are also most probably connected to each other.

Within the study of small-world on social networks and more generally complex networks, some other properties of these networks have been observed:

- Power-law degree distribution or preferential attachment or scale-free property is theoretically representing the well-known behaviour that "rich gets richer". From the perspective of graph theory, in a network represented by a graph, vertices who have more connections are likely to get even more new connection in future. [7]

- Sparsity in a graph, refers to the relative number of edges to the number of vertices. Graphs associated with social networks are sparse. Thus, each vertex is connected to a small ratio of vertices in the graph.

- High clustering coefficient is another property of social networks. As an example in a friendship network high clustering coefficient indicates the fact that friends of a person are most probably friends with each other as well.

2.3 Metrics and Measures

In this part we summarize measures and metrics which are used in social network analysis:

2.3.1 General Metrics for All Networks

Considering a network as a connected graph or the giant connected component of a graph the following properties are defined:

- Density: the number of edges in the network to the number of possible edges. Mathematically, it is defined as $\frac{2E}{N(N-1)}$ where E and N denote the number of edges and nodes respectively.

- Size: most of the times size of a network refers to the number of nodes N, but less likely the number of edges E which takes a value between $(N-1)$ and $\frac{N(N-1)}{2}$.

- Average degree: degree of a node in a network is the number of its connections. Thus, average degree is be computed by $\frac{2E}{N}$.

- Average path length: refers to the average shortest path length, the minimum number of edges two nodes are away from each other, between all pairs of the nodes in a network. Hence, it shows how close on average two random nodes are to eachother.

- Diameter: is the maximum shortest path between all pairs of nodes in a network. Diameter of a network is a measure that shows the upper bound of how far nodes are in that network.

- Clustering coefficient: shows how well neighbours of a node are connected to each other. Clustering coefficient for a node i is computed as $\frac{2e_i}{k_i(k_i-1)}$,

where k_i is number of neighbours of node i and e_i is the number of edges between them. If the neighbours of a node are making a clique, the clustering coefficient of that node would be 1, and if they are not connected at all it is 0. The clustering coefficient of a network is the average over all nodes.

2.3.2 Centrality Measures

In complex networks such as social networks, we are interested in ranking nodes and finding important nodes. To this end, centrality measures have been introduced to help us calculate how much a node is central (centre of importance) according to some criteria. Here, we present the most well-known centrality measure as follows:

- Degree centrality: ranks nodes based on their degree. It is calculated by the normalization number of edges at each node $C_d(v_i) = \frac{\sum_{j=1}^{n} e(v_i,v_j)}{N-1}$ [27], where $e(u, v)$ is 1 if there is an edge between u and v.

- Closeness centrality: shows how much a node is close to other nodes in the network. It counts the average number of hops a node is away from the rest of the graph. Thus if the average distances of a specific node with the rest of the graph is small, the closeness centrality of that node is high and vice versa. Hence, closeness centrality has an inverse relation with distances in the network and is calculated by $C_c(v_i)^{-1} = \frac{\sum_{j=1}^{n} d(v_i,v_j)}{N-1}$ [27], where $d(u, v)$ is the distance between u and v.

- Betweenness centrality: shows the importance of a node in controlling the communication between other pairs of nodes in a network. In order to measure this characteristic, betweenness centrality is the count of the number of shortest paths a node lies on and is calculated by $C_B(v_k) = \frac{2\sum_{i}^{n}\sum_{j<i}^{n} b_{ij}(v_k)}{N^2-3N+2}$, where $b_{ij}(v_k)$ is the probability of v_k to be on the shortest path between v_i and v_j [27].

- Eigenvector centrality: is based on the idea that a node is more central if it is connected to central nodes. Therefore, in addition to neighbours of a node, their centrality value is taken into account. Formulating this concept

forms a well-known eigenvalue, eigenvector equation [58]. The concept of eigenvector centrality was first introduced in [11].

Figure 2.1 shows how these centrality measures differ in ranking nodes in a network. More centrality measures are defined in the literature such as Katz centrality [33], Alpha centrality [12], etc., in addition to the aforementioned ones. We do not introduce other centrality measure as they are beyond the scope of this dissertation.

Centrality measures are all node-centric as they are defined for individuals in a network. Freeman in [27] has extended the previously defined centrality measures for the whole graph. According to [27], centrality of a graph is a measure to show the tendency of a node to become more central. Graph-centric centralities are defined as follows:

$C_X(v_i)$: one of the node centralities defined above.
$C_X(v^*)$: largest value of the $C_X(v_i)$ for any point in the network.
$max \sum_{i=1}^{N}[C_X(v^*) - C_X(v_i)]$: the maximum possible sum of differences in node centrality for the entire network consisting of N nodes. Then C_X calculates centrality value for the whole graph as below:

$$C_X = \frac{\sum_{i=1}^{n}[C_X(p^*) - C_X(v_i)]}{max \sum_{i=1}^{N}[C_X(p^*) - C_X(v_i)]}$$

Centrality measures in general, are used to rank nodes in a network. In addition to centrality measures, there exists other ranking algorithms for ranking nodes based on a score computed for each node. Among these ranking algorithms, two well-known ones are as follows:

- HITS [16]: Hyperlink-induced Topic Search which is also known as *hubs and authorities* is a link analysis based rating algorithm for Webpages. They introduce two kinds of pages in [16]: hubs and authorities. Hubs are pages containing many links to other pages and authorities are pages having a good content. Of course a good hub is the one that contains more links and a good authority is the one that is linked from many hubs. HITS algorithm assigns

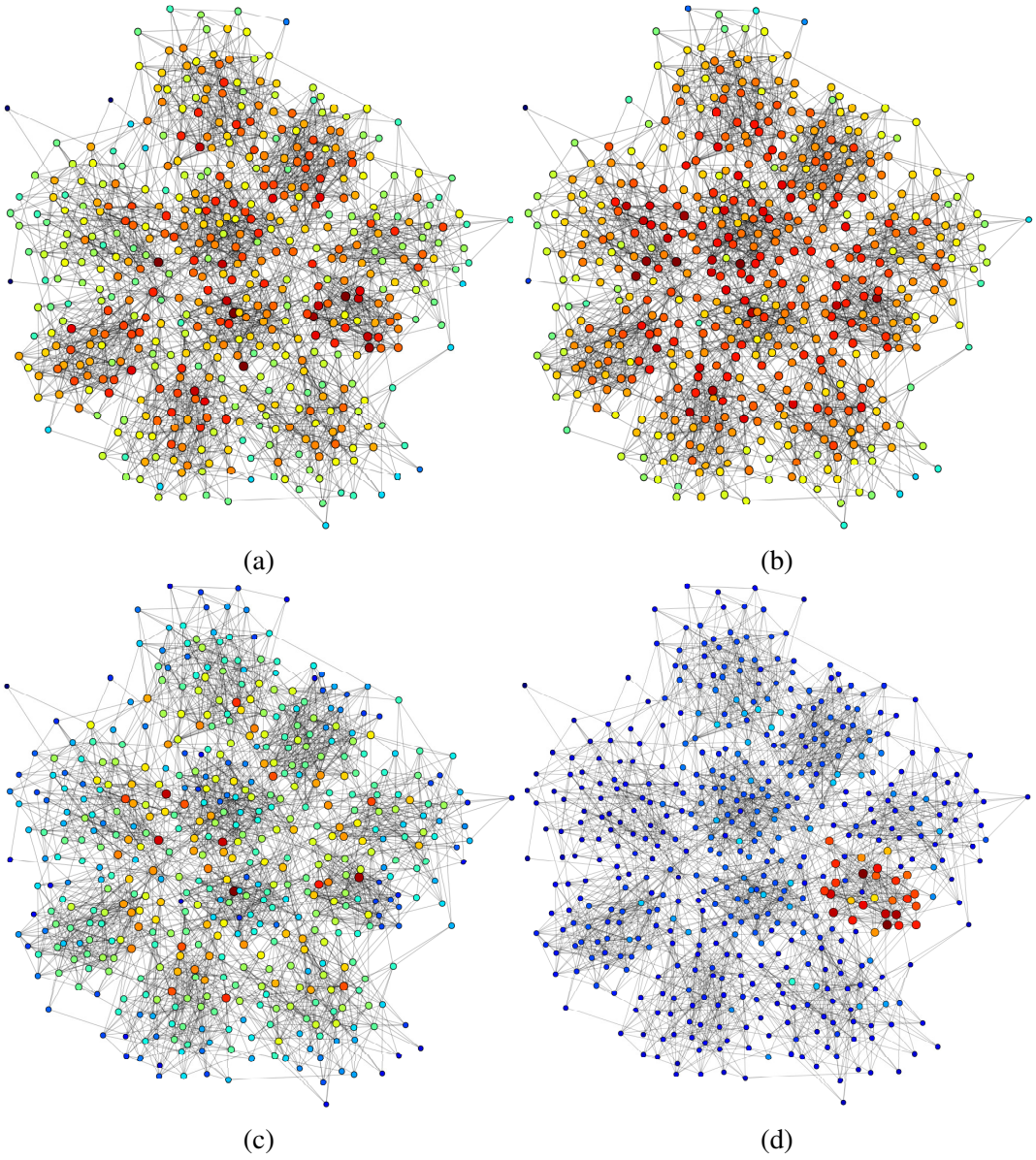


Figure 2.1: Central nodes based on different centrality measures in the network formed by members of a karate club. Colors are changing from dark red to dark blue. The more red the color of a node, the more central it is. Similarly, size of a node is also an indicator of its centrality score. The larger the node, the higher is its centrality score. (a) Degree centrality, (b) Closeness centrality, (c) Betweenness centrality, (d) Eigenvector centrality.

hub and authority scores to each page. These two scores are iteratively updated a fixed number of steps or until convergence.

- Pagerank [54]: The most famous algorithm based on eigenvector centrality is *Pagerank*. Consider the scenario that a user starts clicking randomly on links in Webpages. The interesting question is how likely it will end up in a particular Webpage. Pagerank shows the probability distribution of the likelihood of ending up in each Webpage in the above scenario. Pagerank is the first algorithm which is used to rank Webpages for Web search by Google.

These ranking algorithms were originally developed for Web graphs, but they are also used as measures to rank nodes in social network graphs.

2.4 Dynamic Social Network Analysis

According to some believes, time is the fourth dimension of this world [28]. In truth, It is an inseparable characteristic of all systems and phenomena happening in this world. Considering time in modeling and formalizing systems results in a much more complicated problems. Nevertheless, some concepts, assumptions, and reasonings may completely change or be violated when this fourth dimension comes into account. Notably, social networks are no exception considering the effects of time. Social networks are intrinsically temporal since they are made up of humans. Considering the fact that people change their relationships, affiliations, hometown, job, etc. through their lives, the formed societies also changes accordingly. Altogether, the dynamicity of social networks is accepted worldwide, however, models, metrics, algorithms, and analysis are mainly focused on a static snapshot of a network [60].

Models considering the effect of time in a network use the term *temporal*. Kostakos presents the idea of temporal graphs as means for modeling networks in [42]. In this work by Kostakos, graph definition, construction methods and a number of metrics are introduced to be used for exploring temporal behaviours of a network. Kostakos constructs the temporal graph as follows:

1. For each entity, create a node in each timeframe if that entity is interacting with others during that timeframe. This way, there may exist multiple nodes representing the same entity but in different timeframes. For instance, if entity A is interacting within timeframes $t_{i_1}, t_{i_2}, \cdots, t_{i_k}$, the set of nodes $\{A_{t_{i_1}}, A_{t_{i_2}}, \cdots, A_{t_{i_k}}\}$ represent nodes associated with the entity A in those timeframes.

2. Add a weighted directed link between all consecutive time versions of an entity as $(A_{t_{i_x}}, A_{t_{i_{x+1}}})$ and set the weight to be $t_{i_{x+1}} - t_{i_x}$.

3. For each interaction add an unweighted (directed) link between entities that interact with each other in each timeframe.

Without loss of generality, the third rule presented in [42] could be generalized as follows:

- Based on properties of the underlying network, add a (un)weighted (un)directed link between pairs of interacting entities in each time frame.

In the model presented by Kostakos, the only links between timeframes are the links between time versions of an entity. In other worlds, latency on interactions between distinct entities is not considered. Adding this functionality to the model makes it more powerful and complicated at the same time. With this structure for the graph, Kostakos [42] defines temporal metrics that help getting more insight about the network:

- **Temporal proximity** $p(X, Y, t_a, t_b)$**:** measures the time it takes to go from X to Y when starting in X at t_a and reaching Y at t_b. t_a and t_b can either be a value or *null*. Based on values for t_a and t_b, temporal proximity can be computed in 4 ways:
  - $p(X, Y, t_a, t_b)$: the temporal shortest path between X at t_a and Y at t_b.
  - $p(X, Y, t_a, null)$: the temporal shortest path considering X at t_a and any version of Y.

- $p(X, Y, null, t_b)$: the temporal shortest path considering X at any time and Y at t_b.
- $p(X, Y, null, null)$: the temporal shortest path considering any version of X and any version of Y.

weight of a temporal path is calculated by summing over all temporal weights on links.

- **Average temporal proximity:** measures on average the time it takes to go from X to Y and is defined as

$$P(X, Y) = \sum_{i \in \mathcal{T}} p(X, Y, t_i, null)/N, \quad p(X, Y, t_i, null) \neq null$$

Based on P, two other average temporal proximity measures P_{in} and P_{out} are also defined as

$$P_{in}(X) = \sum_{i \neq X} P(i, X)/N, \quad P(i, X) \neq null$$

$$P_{out}(X) = \sum_{i \neq X} p(X, i)/N, \quad P(X, i) \neq null$$

$P_{in}(X)$ shows how fast, in terms of time, X is reached by the rest of the network, and similarly $P_{out}(X)$ measures how fast X can reach the rest of the network.

- **Geodesic proximity** $g(X, Y, t_a, t_b)$**:** measures the edge distance between X at time t_a and Y at time t_b as the number of hops between them. Similar to temporal proximity, geodesic proximity can also be computed in 4 different ways according to the values of t_a and t_b.

- **Average geodesic proximity:** measures the average number of hops X is away from Y and is computed as

$$G(X, Y) = \sum g(X, Y, t_i, null)/N, \quad g(X, Y, t_i, null) \neq null$$

Similar to the average temporal proximity, G_{in} and G_{out} are defined as the average number of hops X is reached by the rest of the graph and the average

number of hops does X reach the rest of the graph respectively:

$$G_{in}(X) = \sum_{i \neq X} G(i, X)/N, \quad G(i, X) \neq null$$

$$G_{out}(X) = \sum_{i \neq X} G(X, i)/N, \quad G(X, i) \neq null$$

- **Temporal availability:** is a measure that shows the probability of existence of a path between two nodes at any given time and is defined as:

$$V(X, Y) = |\{g(X, Y, t_i, null) \neq null\}|/N$$

Similar to above metrics, two other availability metrics V_{in} and V_{out} are defined as:

$$V_{in}(X) = \sum_{i \neq X} V(i, X)/N \quad V(i, X) \neq null$$

$$V_{out}(X) = \sum_{i \neq X} V(X, i)/N \quad V(X, i) \neq null$$

Casteigts et al. in [14] integrate the existing models for studying the dynamicity of networks into a unified framework called TVG (time varying graphs). Given a set of entities as vertices(V), a set of relations between these entities as edges(E), and a set of labels(L) for edges, time varying graph, TVG, $\mathcal{G} = (V, E, L, \mathcal{T}, \rho, \zeta, \psi)$ is defined. Where $\mathcal{T} \subseteq \mathbb{T}$ is the lifetime of the network, $\rho : E \times \mathcal{T} \rightarrow \{0, 1\}$, is the presence function indicating which edge is present at each time step. The next parameter, $\zeta : E \times \mathcal{T} \rightarrow \mathbb{T}$, is latency function that shows the time it takes to cross a given edge strating at a given time. And finally $\psi : V \times \mathcal{T} \rightarrow \{0, 1\}$ is the presence function for nodes. The TVG framework just described can be used to describe networks from transportation, communication, social, and more generally complex networks.

Santoro et al. in [60], classify social network indicators into temporal and atemporal. Temporal indicators are defined through time on consecutive timeframes, whereas atemporal indicators are static characteristics defined in a single timeframe. They further study the evolution of these indicators based on TVG formalism defined in [14].

In a dynamic network, where time is considered, the definition of paths between entities become tricky. A Path, is a fundamental concept in social network analysis as many metrics, measures, and algorithms are based on that. When it comes to time-varying graphs a *journey* is a fundamental concept that is an extension of the notion of path [60]. Xuan et al. in [81], define the concept of journey and develop distance measures based on journeys instead of paths. *Hop-count* or length of a journey denotes the number of nodes on the temporal-path. *Arrival date* indicates the time we reach to the end node of the path which is sum of the scheduled time to start the last edge and the time it takes to traverse it. Finally, *journey time* defines the time between starting of the temporal-path and the arrival. Based on the concepts of hop-count, arrival date, and journey time three important types of journeys are defined in [60] as follows:

- The *distance* in a graph between two nodes u and v is defined as the minimum hop-counts taken over all journeys between u and v. This journey is called *shortest*.

- The *earliest arrival date* in a graph between two nodes u and v is defined as the time of the first journey arriving at v from u. This journey is called the *foremost*.

- The *delay* in a graph between two nodes u and v is defined as the minimum journey time, taken over all journeys in the graph between u and v. This journey is called *fastest*.

Based on these three definition of distance in temporal graphs, multiple versions of diameter, centrality measures and other metrics, that are based on paths between nodes, could be defined for temporal graphs. Kossinets et al. in [41] show that nodes that are topologically central are not necessarily central when time in considered. Thus, there is a need to redefine centrality measures for temporal networks. In this regard, temporal betweenness and closeness metrics are recently introduced in [71].

In our proposed framework, roles are defined in a single timeframe of the network. Thus, the set of roles defined in the structural social role mining are basically

static. However, using temporal centrality measures introduced in this section results in identification of dynamic (temporal) roles.

2.5 Community Mining in Social Networks

Social networks are sparse graphs compound of dense subgraphs. These dense subgraphs are called *communities*. The most consensual definition reached on communities defines a community as a group of nodes in a graph which are highly connected to each other and have less edges to the rest of the network. Some reachers believe that community detection is a version of graph partitioning, hence they are looking for cuts to identify communities. While some other researchers believe that overlapping should be considered in community mining, thus it is not similar to cutting graph into parts. There are many works on community extraction and identification which can be categorizes in different ways. Here, we only present prominent algorithms from the literature.

Community mining algorithms are based on links between nodes that indicate the connectivity of two entities. SCAN (Structural Clustering Algorithm for Networks) [80] is a method for detecting communities according to how nodes are sharing their neighbours in addition to only considering direct connections. Thus, if two nodes are connected and are also sharing a reasonable number of their neighbours, they belong to the same community.

Palla et al. in [55] define community as a subgraph that consists of a set of complete subgraphs sharing many of their nodes. More formally, a k-clique community is a union of all k-cliques that can be reached by eachother through a series of adjacent k-cliques. By the notion of adjacency they mean sharing $k-1$ nodes between two k-cliques. This method is among the first community mining algorithms which allows overlapping between communities.

Having the intuition that random walks are most probable to get trapped in highly connected parts of a dense graph, Pon et al. propose a community mining algorithm [56] using random walk to compute nodes' similarities. The node similarity measure is used to partition nodes in different communities in the graph.

Rabbany et al. in Topleader [57] propose a new algorithm for community mining based on the intuition that in each community there exists a set of leaders and other members of that community are followers of the leaders. In their proposed algorithm, they set k initial leaders and find their followers to form the initial communities. Afterwards, they update leaders in each community, and consecutively update followers and so forth until convergence.

Newman in [50] develops the fast modularity algorithm for detecting communities in a network. The invariant, modularity is defined as $Q = \sum_i (e_{ii} - a_i^2)$ where e_{ii} is the fraction of edges in community i and a_i^2 is the fraction of edges having one end in community i when the other end which is out of the community is randomly connected to the other end of similar edges. In fast modularity algorithm, each node is assumed to be a community initially and communities are joined in an order that maximizes the increase on the modularity (Q) value.

Dynamic Community Mining in Social Networks

Considering temporal characteristics of social networks, detecting communities becomes a more challenging problem. Due to the fact that, in different snapshots the structure communities may change in terms of entities and their connections. The question arising here is how to identify two equivalent communities from different snapshots. Tantipathananandh et al. in [72] propose frameworks and algorithms to study dynamicity of communities through time. They reduce the problem of community matching to a graph coloring problem in their paper.

Takaffoli et al. in [70, 69] propose the MODEC framework to study the evolution of communities in different timeframes of the network. In this regard, they define a series of events important in a community's life cyle: *form*, *survive*, *split*, *merge*, and *dissolve*. These events respectively refer to the appearance of a community, a community continuing its life, a community's decomposition into multiple communities, multiple communities becoming one larger one, and a community disappearing in a timeframe. They also propose a matching algorithm to find community matches in different timeframes. Furthermore, they define the concept of

meta-community that refers to all instances of a community from different time-frames.

2.6 Social Media

Social networking websites, media sharing application and blogs have been growing in their usage and popularity among individuals and businesses. The capability of creating and sharing content and the speed the contents can propagate on these applications, are keys to make them powerful. Many of us may think of friendship networks when we hear the word social network, however, they are more than that. People are using these platforms to create content, share news, talk about their ideas, advertise for products, discuss political, economical, and social issues. Facebook[1] and Twitter[2] are two well-known examples of these webpages. With all these capabilities and usages, social networking platforms act as a real media and in some aspects they are even more powerful than other forms of media.

There are many works on applications of social networks such as viral marketing [22, 44], revenue maximization [31, 3], influence [36, 15, 2, 4] and many other topics. The extensive number of research in these areas, shows how impactful they are. Kwak et al. in [43] investigate characteristics of tweets and posts from Twitter and compare contents of the posts to news headlines. According to their findings, many breaking news appear in Twitter ahead of CNN. This phenomena is caused by the live broadcasting nature of the tweets. In another work by Sakaki et al. [59], the authors develop a system based on tweets that detects earthquakes in Japan and sends notification to a set of registered users. The developed system, which is based on Twitter posts is reported to send notifications faster than broadcasts by JMA (Japan Meteorological Agency).

On the other hand social networking platforms provide users with tools to make their voices be heard. According to a BBC article [76], "These days, one witty tweet, one clever blog post, one devastating video, forwarded to hundreds of friends at the click of a mouse, can snowball and kill a product or damage a company's

[1] http://www.facebook.com
[2] http://www.twitter.com

share price". In this regard, Kietzmann et al. in [38] emphasize on the importance of social media for businesses. They provide a framework based on 7 functional building blocks (identity, conversations, sharing, presence, relationships, reputation, and groups) to help businesses understand social media. In another work by Kaplan et al. [32], the importance of social media is being highlighted. According to the authors in [32], although social media is not well understood, it is the agenda for many businesses. Thus, decision makers and consultants try to find ways to make profitable use of these applications. They discuss challenges and opportunities businesses have regarding the fast evolution of social media. Moreover, they provide a set of recommendations for companies to build their social media strategies.

In approval of the importance of social media, we refer to [5] by Asur and Huberman. In this work, they show how social media can be used to predict future. In their experiments, they use information from Twitter to predict box-office revenues for movies. They developed a linear regression model for their predictions and show that accuracy of their results outperform those of Hollywood Stock Exchange.

Our proposed role mining framework, provides tools to analyze and understand social media. The extensive information available from the social media might seem confusing for businesses to use and build their strategies upon. However, identifying roles could help in getting less but more valuable information out of that.

2.7 Roles in Online Social Networks

2.7.1 Roles Theory

Role is a fundamental concept in social sciences. Many frameworks for studying various social issues use the concept of role. Although it is an important concept, there is no consensus on the definition of it. Thus, what the concept "role" means is still a question. Biddle integrated sporadic works to define the concept of role by his work in [10] and defined the field of role theory.

Role theory studies the concept of role and integrates various models and the-

ories about role. According to the role theory, the concept of role is explained by assuming people as members of social positions that have expectations of themselves and others' behaviours.

In [10], various theories regarding the concept of the role are integrated which we briefly presented in this section. None of these theories are complete and ideal and each one is focuses on some aspects and misses some other points.

- *Functional role theory* focuses on characteristic behaviors of people that have a social position in a stable social system. "Roles" are the shared normal expectations that explain these behaviours. Functional role theory does not consider that many roles are not associated with identified social positions and social systems are far from stability.

- *Structural role theory* focuses on *social structure* as stable organizations of sets of people (called *social positions* or *statuses*). These people share the same patterned behaviours (*roles*) that are directed towards other sets of people in the structure. The focus of structural role theory is more on the social environment and less on individuals. Moreover, the arguments in structural role theory are more likely to be presented in mathematical models compared to all other theories on role.

- *Organizational role theory* focuses on social systems that are preplanned, task-oriented, and hierarchical. Roles in such organizations are assumed to be associated with identified social positions and to be generated by normative expectations. However, these norms may vary among individuals and may reflect both the official demands of organizations and the pressures of informal groups.

- *Cognitive role theory* focuses on relationships between role expectations and behaviours. This theory is more concerned about social conditions that give rise to expectations, techniques for measuring expectations, and the impact of expectations on social behaviours. According to [10] cognitive role theory ignores dynamic characteristics of human interaction. As well, cognitive

role theorists often underestimate role phenomena associated with social positions or with temporal and structural phenomena by focusing too much on individuals.

A Mathematical Model for Structural Role Theory

One big criticism to role theory is the lack of mathematical modeling for the concepts within the theories. Oeser and Harary have built a mathematical framework [53, 51, 52] which sees a role as a structural concept, for modeling the *structural role systems*. In this part, we introduce their model in detail.

Elements of a role structure:

- Task: Every group or community is formed to achieve a set of goals. Thus, a set of jobs are defined to fulfill those goals in each community. This set of jobs as a whole is divided in elements called tasks or task elements which are defined in time and place among members of the group.

- Position: It is a title in social structure which can be occupied by individuals meeting necessary requirements. These requirements come from a set of rules that define the abilities required for the task associated with each specific title. So titles determine tasks of people. *Position* is defined in terms of:

  - sociological characteristics: such as education,
  - psychological characteristics: such as skills, intelligence, leadership abilities,
  - relationships to other people or positions,
  - ability to perform certain tasks.

  Hence through the definition of a position, we would be able to identify special types of individuals, their duties toward others in different positions and their part in the task system of communities. In brief, task and people are real elements, whereas position is a concept defining sets of relationships between tasks and people.

- Person: A “person” (a set of attributes) is a human being who is related to task elements through a set of relationships called positions. Thus, a “person” who is assigned to a certain position has no attributes other than those defined in the specification of tasks of that position. This definition of the concept of person in the task system is based on the rule that everyone is replaceable in the system.

The concept of role indicates a position that includes persons, positions and tasks. Therefore, the necessary elements to define a structural role system are persons, positions and tasks. To complete the model, the set of relationships between the elements are:

- Rules that define the selection of persons for positions.
- Rules that define the relationships between positions.
- Rules that allocate tasks to positions.

Obviously, there might also exit informal relationships, however, in role theory the focus is on formal relations and informal ones and their impacts are not considered. Structural role system is mathematically defined as below:

- H-graph: Original nominal scale that shows the set of people.
- P-graph: It is an organizational chart of the institution defining the hierarchical relations on the set P of positions. In this graph, nodes are positions and directed edges between nodes indicate the power relationships (starting from the boss to secretary).
- T-graph: Is the graph that shows the work layout. Nodes represent tasks and directed edges indicate relationships between tasks. Hence, this graph shows the order in which tasks are to be done.
- H-P graph: This graph consists of H nodes from H-graph and P nodes from P-graph as well as the links between them that show who is assigned to each position.

- P-T graph: This is the task allocation graph showing which position is responsible to do which task(s).

Definition of the formal role of position P

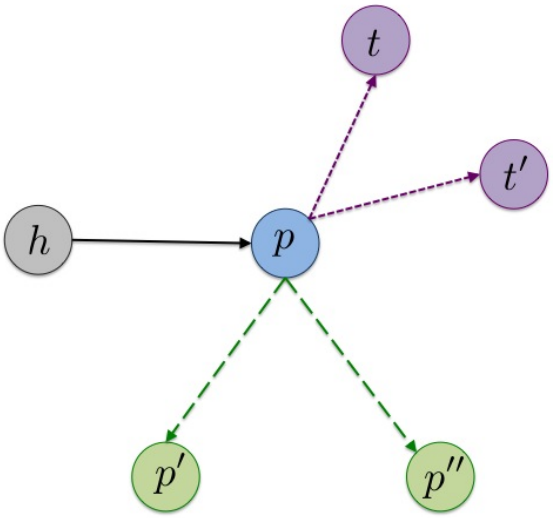


Figure 2.2: Directed graph for defining the role of the position p from [10].

The directed graph for defining a role is called D which is shown in 2.2 and describes definition of the formal role of the position p. Assume the relationships from the point of view of position p in Figure 2.2. The components of a role are defined as follows:

- Person h who is assigned to position p.
- Position p that has immediate sub-positions p' and p''
- Tasks t and t' that are assigned to position p.

In the structural role system the important characteristic is that tasks are not directly assigned to individuals, there are assigned to positions and positions to individuals. In the structural role graph D, the *job* of a *person* h is defined as all task elements that are reachable from h in that graph.

Within a social network we are not usually provided with the information about positions, tasks, and person. We only have individuals, their connections, and sometimes extra attributes that can give us some information about these three components of a role. Thus, we have to come up with ideas for defining role in network science using the definitions from sociology.

2.7.2 Definitions

In this section, we present some of the proposed definitions of the concept "role" in the literature.

- "A node role is a subjective characterization of the part it plays in a network structure." [62]

- "Individuals' behavior in groups is constrained by several factors, including the skills, privileges and responsibilities they enjoy. We call these factors a social role."[30]

- "Peoples behavior in social situations is not random and completely unpredictable, nor is it uniformly identical in each situation. Rather, people act differently toward different people, and depending on the circumstances at hand; this much is readily apparent. The reason is that, besides having personalities, by being part of a social group, people occupy positions in the social structures of groups that allow them to do and say certain things, as well as constrain them from saying and doing other things. This mixture of allowances and constraints, combined with the choices the individual makes given this mixture, constitutes a social role." [30]

- "A behavioral repertoire characteristic of a person or a position." [9] which describes both formal and informal roles. Formal roles are the ones where the behavioral characteristic of the indivdual is the result of the role, whereas in informal roles, the role is recognized because of a set of behavior an individual has. [30]

- "Rights and duties attached to a given status." [29] This definition is important as it shows that roles are coupled with reponsibilities. [30]

The most prominent part in the definition of role, which should not be forgotten, is that roles are defined through interactions between individuals. Roles are a set of behaviours and statuses along with a set of responsibilities that are meaningful merely through interactions between individuals. For instance, a graduate student is performing the student role in an interaction with his supervisor. However, the same student plays a teacher or instructor role when he, as a teaching assistant, interacts with undergraduate student. For this example, we can see that roles are relative to the context and relationships between individuals.

2.7.3 Roles in Social Network Analysis

Forestier et al. in [24] have done a thorough survey in 2012 on 'roles' in social networks. According to their survey, roles are categorized as non-explicit and explicit ones. Non-explicit roles are identified in an unsupervised framework, which requires little information about the roles beforehand. Thus, clustering algorithms can be used to identify them by using structural or contextual information as features in a network. Whereas, explicit roles are defined as specific measures beforehand and are identified by calculating those measures for each node in the network.

Non-Explicit vs. Explicit Roles

Blockmodeling is an example of methods for identifying *non-explicit roles*. It is an algebraic framework to cluster nodes in a network mainly based on the structure of the network. Bockmodeling has various application in social networks including identifying roles. Doreian et al. in their book [23] overview the usage of blockmodeling in social networks. A blockmodel is the adjacency matrix of a network that shows the relations between nodes. An example of a blockmodeling process is shown in Figure 2.3 [24]. The matrix in Figure 2.3(b) represents the relations between 8 nodes in the network in Figure 2.3(a). The adjacency matrix is transformed to a matrix shown in Figure 2.3(c) by a set of permutations on columns and

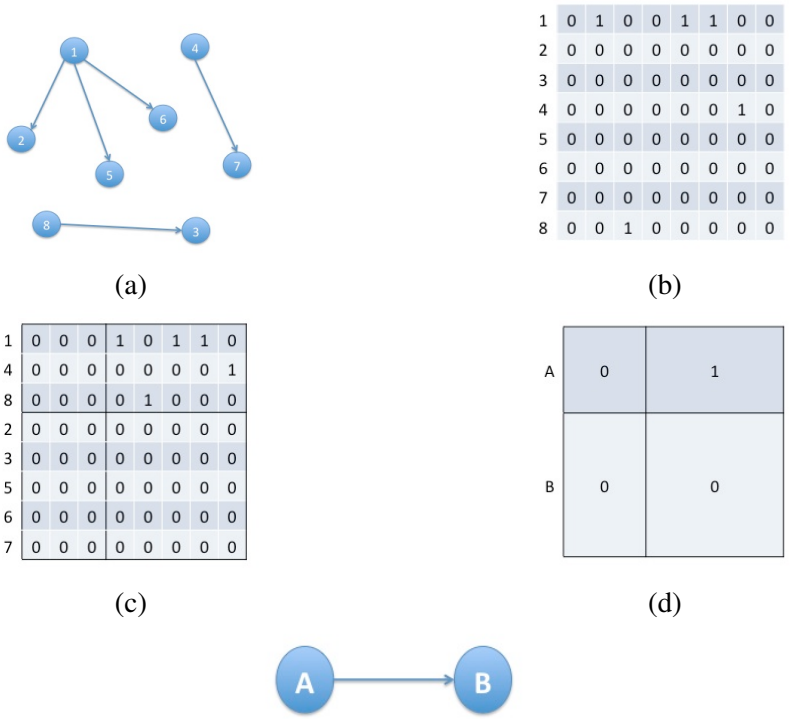


Figure 2.3: The process of blockmodeling from [24].

rows. Permutations can be done in both supervised or unsupervised fashion. Figure 2.3(c) shows four blocks formed in the permuted matrix. Three *zeroblocks* and a *non-zeroblock* containing some 1s. According to the used equivalence notation, the matrix in Figure 2.3(c) can lead to the block model represented in Figure 2.3(d) representing two positions shown in Figure 2.3(e).

For the *explicit roles*, which are the predefined roles, the identification process is through satisfaction of predefined criteria by some nodes in the network. Two well-known explicit roles are *experts* and *influentials*.

definition: *An* expert *is a person who is knowledgable about a topic and is skillful in that specific topic, which makes him trustworthy in the network.*

To see the importance of identifying the role "expert", assume that we are looking for an answer to a question in a technical forum. Identifying experts there helps to find trustworthy answers. Zhang et al. in [83], work on a java technical forum to identify individuals having the expert role. They construct the network by con-

sidering directed links from a user who has asked a question and those who have answered that question. They propose three algorithms based on z-score, pagerank and HITS using indegree and outdegree of nodes to identify experts. They have also compared results on simulated networks and found out that the structure of the network has a significant impact on the results of ranking experts.

In addition, expertise is relative to the area or topic. An individual who has a great level of expertise in politics might not necessarily have a high level of expertise on programming. Balog and De Rijke [6] consider expertiseness on a specific topic by defining a topical profile for each user. The topical profile of each user presents the probability of the user being expert in several particular topics. Each person is considered relevant to a topic if he is mentioned or has a relevant document in that topic.

In [21], the focus is on extracting experts on a specific topic in the network of emails. The expertiseness is computed pairwise for individuals. Thus, it is relative between two individuals who have responded to each others' emails. The authors have used various measurements and they show that PageRank outperforms other algorithms for identifying experts.

Identifying *Influentials* or *Influencers* has also attracted great attention from researches as it has many significant applications in viral marketing, spread of information and innovation, etc.

definition: *An* **influencer** *is a person who has the capability or power to influence the decisions, thoughts, actions, behaviors, etc., of other people inside a social network.*

> *"The Influentials*
>
> *Who are they? The most influential Americans – the ones who tell their neighbors what to buy, which politicians to support, and where to vacation – are not necessarily the people you'd expect. They're not America's most affluent 10 percent or best-educated 10 percent. They're not the* early adopters *always the first to try everything from Franco-Polynesian fusion cooking to digital cameras. They are, however, the 10 percent of Americans most*

> *engaged in their local communities $\cdots$ and they wield a huge amount of influence within those communities. They're the campaigners for open-space initiatives. They're church vestrymen and friends of the local public library. They're the Influentials $\cdots$ and whether or not they are familiar to you, they're very well known to the researchers. For decades, these researchers have been on a quest for marketing's holy grail: that elusive but supremely powerful channel known as word of mouth. What they've learned is that even more important than the* word *– what is said – is the* mouth *– who says it. "* [34]

In [39], Kim and Han have interpreted influentials in three directions: The ones with characteristics more like a *Sales Person*, which is more reasonable in the context of selling and buying a product. The ones with characteristics more like *Opinion Leaders*, which makes more sense in the context of political elections, adapting an idea, etc. Finally, the ones with characteristics more like a *Connector*, who are important ones in bringing people together and probably building new communities. Kim and Han in their work [39] have considered influential people who are more like sales persons. They have developed a two-step methodology for identifying influentials by means of the structural properties of the network. In the first step, they identify potential influentials based on degree centralities. In the next step, the set of potential influencers are analyzed by their activity history to identify influential people.

Kempe et al. have studied the problem of identifying a set of influentials in a network that lead to the maximum diffusion of innovation in the network [37]. Their work is based on the observation that individuals' decisions to adopt an idea or buy a product are directly affected by their friends and acquaintances. They study the propagation of influence in the decreasing cascade model in a network. In this model, influence is propagated in a cascading process defined by a probabilistic rule. Thus, they begin with a set of initial active nodes and study the size of the target set. In their work, they have proved an approximation factor of $1 - 1/e - \epsilon$ for their results compared to optimal solution if an initial set of size k is selected in

the natural greedy algorithm.

Diffusion of innovation is another direction in which detecting influential nodes is important. The theory of diffusion of innovation has been formed in anthropology, sociology and epidemiology. This theory explains how opinions and products may spread in a community. Valente and Davis in [74] show that the diffusion process increases when initiated by opinion leaders. Opinion leaders are those who are acting as role models in a community. There are some studies showing the importance of opinion leaders in different issues such as decreasing the rate of unsafe sexual practices [35] and decreasing the rate of cesarean births [45]. Moreover, in [74], Valente and Davis show that maximizing the effectiveness of opinion leaders lead to a faster rate of diffusion. In their work, opinion leaders are identified based on direct ties. People nominate leaders and those who get nominated more are chosen as opinion leaders.

In another work by Agarwal et al. in [2], influential individuals are identified in the context of blog posts. They propose a novel method for identifying influential bloggers in Blogosphere based on identifying influential blog posts. An *iIndex* is calculated for each blogger based on his influential blog posts. Moreover, they introduce four metrics representing the influence of a blog post.

- **Recognition:** It is how much an influential blog post is referred by other posts. Thus, *the indegree of a post* (ι) shows the number of times it has been referenced by other posts.

- **Activity Generation:** The capability of a blog post in attracting others' attention is a measure of how influential it is. *Number of comments* (γ) on a blog post shows how much it is successful in attracting others' attention to some extent.

- **Novelty:** Novel ideas attract more attention than repetitive ones. Therefore, novelty is another characteristic of an influential post. The number of references in a post, which is *the outdegree or outlinks of that post* (θ) is negatively correlated with novelty. To be more elaborate, the more references a post has to other posts shows that it is using contents and ideas from those posts. Thus,

the number of references in a post is an indicator that the post is not novel.

- **Eloquence:** An influential post is well-written as the language is the fundamental tool of presenting a concept to others. In blogs environment, it is not reasonable to write long posts, as audiece will not read long posts so often. However, when a post is long, it is a good evidence of its quality. If the author is not confident with the quality of the content, he would not risk loosing audiences by writing that content in a lengthy post. Thus, *the length of a post* (λ) which positively relates to the number of comments, is a heuristic to show the influence of it.

In [2], each blog post gets an *Influence Score (I)* which is calculated with respect to the four properties defined above (θ, λ, γ *and* ι). This score is used to define influential bloggers. An Influential blogger is explained to be an individual with at least one influential post. Therefore, an *iIndex* is introduced for each blogger B as follows: $iIndex(B) = max(I(p_i))$ where p_i is the i^{th} post of blogger B. In addition, the authors have also studied the changes of *iIndex* of bloggers through time and based on their findings, they categorize influential bloggers into four classes:

- **Long-term influentials** Who are an influential for a very long time and can be considered as authorities in the community.
- **Average-term influentials** Who stay as an influential for a shorter period of time (4-5 months based on their data in [2].
- **Transient-term influentials** Who are influentials for a very short time period (1 or 2 months).
- **Burgeoning influentials** Who are becoming an influential recently.

According to this classification in [2], long-term influentials are more affective and more trustworthy for targeting.

Agarwal and Liu in [1] define influential as an individual who is prominent in diffusion of influence. Hence, influentials are different from initiators of an idea or creators of a content. They are more important because of their position in the

network. Simply, influentials not only have to come up with novel ideas, but also they should have strategic position in the network to diffuse the idea.

Roles in Online Discussion Forums

Golder and Donath in [30] have presented a framework that uses social roles to analyze and explain the behaviours of individuals in an online discussion forum. They have defined a set of social roles for an online newsgroup community, based on frequency of participation, communicative competence and common ground:

- **The Celebrity**

  Celebrities are those who post frequently in the forums and thus are well-known by others such that they could be even the topic that others talk about. Their posts are competence and reliable and in some words, they are the ones who define the community by all their means. They spend much time and effort in their community to make it active and define and protect boundaries. Celebrities are a small percentage of the members of a community, however, they post majority of the posts. The reason why someone should spend this much time and energy in a community can be to get positive self-image or to become famous and could act as a leader to have impact on others.

- **The Newbie**

  Newbies are those who have just joined a community and have less communicative competence and common ground with the group. Thus, they might have a few or no posts until they get involved more.

- **The Lurker**

  Lurkers are the invisible audiences of a discussion forum. They read posts without posting anything in a long time. There might be different reasons for lurkers to not participate such as having fear, not being confident enough, or just not feeling to post anything. On the other hand, lurkers can be of various kinds with different aims as invisible audiences. Since lurkers are often invisible audiences in the community, we face a lack of information to study them.

- **The Flamer, The Troll, and The Ranter**

  These three roles all have negative behavior in a forum.

  - **Flamer:** They have aggressive and hateful language. They do not truly belong to a community and do not follow conversation and goals of the community. They constantly look for victims to engage in a flaming behavior.
  - **Troll:** These individuals pretend to be good members but indeed try to waste communities' conversations and time by their posts. Trolls do not have any specific goal to behave so except for having their joy.
  - **Ranter:** They post as frequent as celebrities and believe in what they say. They have a troll-like behavior to encourage people in their conversations, but they start lengthy, single-minded pointless arguments.

Community-based Roles

In [62], a set of structural roles are defined considering communities in the network: ambassadors, big fish, loners, and bridges. Two parameters are used to identify these roles in a network: degree of nodes, and communities a node belongs to. Figure 2.4 shows the definition of these roles in a plot. As shown in the figure, *ambassadors* are the ones with high degree and also high community metric. This shows that ambassadors are popular nodes who connect various communities together. *Big fish* is important only within a community. Then, the ones with high community score, but low degree, are *bridges* who are more important in connecting communities rather than inside a community. Finally, the ones with low degree and also low community metric are *loners*. In the experiments in [62], it has been shown that in small-world networks, most of the nodes are in the line between loners and ambassadors.

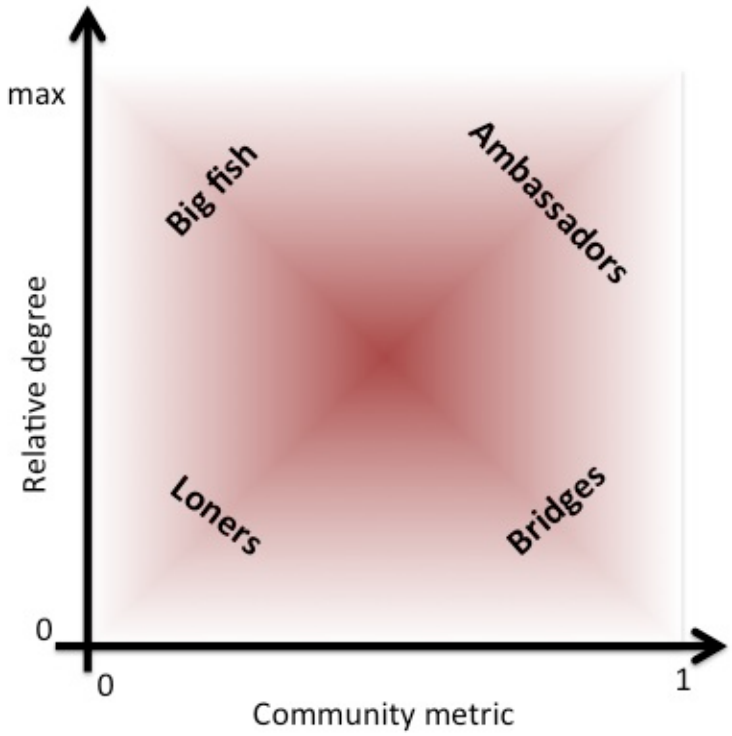


Figure 2.4: Community-Degree chart from [62].

Chapter 3

Structural Social Role Mining Framework

In the desert of life the wise person travels by caravan, while the fool prefers to travel alone.

A middle eastern proverb

3.1 Motivation

Human beings have relied tremendously on their social abilities to overcome difficulties during their long history. In the ancient time, challenges were finding food and water, overcoming threats of wild animals and other tribes' attacks to plunder one's resources, escaping from bad weather, and finding a safe place to reside. Since it was almost impossible to overcome such threats alone, humans formed tribes to live in, hunt, and migrate as groups. This helped them to go through battles of life together and significantly increased their chance to survive.

Although nowadays, we have a safe place to live, we are guaranteed to find enough food in stores, we have technologies to bring water to our houses through central pipelines, and air-conditioning systems help us survive scorching summer days as well as freezing winter nights, more or less we have the same kinds of battles in new shapes. Living in a complicated sociality, cooperation through forming groups and communities is still an important means that not only helps us overcome struggles, but also accelerates the pace of developments.

Groups (societies) are prominent elements of social lives and humans are fundamental elements of groups. The life cycle of a group starts with formation and continues with growth, becoming powerful, weakening, shrinking, and dying as time passes on. Since humans are fundamental components of groups, they play significant roles in groups' life cycles. Obviously, each individual has different characteristics, thus plays different role(s) in the society. For instance, if all students are very shy and introvert in a classroom, friendship groups form very slowly between the students. On the other hand, in the same classroom, if a new extrovert student with a good level of communication skills joins, in a couple of days other students might start building stronger relationships with each other. This is an example showing that influential individuals are fundamental elements of the sociological systems.

In this chapter, we propose a framework to study various roles that individuals may take in a society and how their roles' change through time. The focus of our proposed framework is on structural properties of individuals within their society. More precisely, connections, neighbourhood, and other structural features from the network are used to investigate an individual's role(s). Since the impact of individuals' roles on their surrounding environment is the fundamental reason for the concept of role to be defined, we use this framework to study the effects of roles on phenomena happening in a network. In the example of the classroom, identifying the *social* student helps in understanding why small clusters of students join each other to build a larger cluster of friends. In the following, we discuss our framework in more detail.

3.2 Introduction

With the advent of online social networking applications, many aspects of social life, relations, communications, and cooperations have been reflected in the virtual world of 0s and 1s. For instance, various kinds of digital datasets are now accessible from cooperation networks on Wikipedia, friendship networks on Facebook, academic networks on Academia, organizational relationships from email networks,

news cycles and individuals' reactions and discussions toward them on Twitter, and even more dimensions of human beings' sociality from blogs, technical forums, and customer review forums. The affiliated networks with these datasets can be modeled as graphs in different ways according to the aspect of study. If the concentration is to study the network among people, the dataset can be interpreted as a graph where nodes and edges respectively represent individuals and their relationships. Edges can be directed or undirected, weighted or unweighted depending on the modeling process.

At the first glance, finding influential people in the aforementioned networks may look irrelevant, since the networks are different in many ways. Nevertheless, when modeled as a graph, they share with each other structural properties of a graph. In this work, we focus on the structural properties of social networks' graphs to study role of different members. To this end, we define roles generic to all of these networks based on the structural features of their associated graph. Since the concentration of this study is on social networks, we refer to these roles as "social roles". After defining social roles, we seek for methods to extract them. For this purpose, we suggest a framework called "structural social role mining" where the social roles of the network members are extracted using the structural properties of the network graph.

Our proposed structural social role mining framework is built upon the assumption of having communities as the building blocks of a social network. In the real world, each society is composed of multiple communities and groups. Therefore, people are not just simple entities in social systems, but members of communities and interact with other individuals. In this regard, links between members of a community are not of equal value with inter-community connections. Communities are either explicitly determined in the dataset, or extracted with a community mining algorithm [25].

In reality, each individual may belong to several communities based on various criteria. For instance, we all are members of multiple groups such as family, university, residents of a country, friendship groups, sport teams, etc. Each criterion partitions the network into different combinations of communities. Thus, multiple

criteria lead to several community partitioning in the same network. For example, Figure 3.1 shows a small lunchtime network of a group of graduate students at the University of Alberta. This network, is then partitioned according to 4 different criteria: *gender*, *hometown*, *major* and *degree of their current studies*. As seen, each criterion results in a different partitioning of the network. In this work, we assume having only one partitioning criterion in the studied social networks. Moreover, for simplicity, we assume each node can only be associated with at most one community. Hence, we assume that there is no overlap between communities. However, our framework can be simply extended to the case where communities overlap.

3.3 Definitions

In this section, we review the concept of role in sociology and then bridge between the role concept in sociology and network science. Further, we formally define the social roles that we consider in this study and give the intuition behind considering such social roles.

3.3.1 The Concept of Social Role

Although *role* is a fundamental sociological concept, there is still no consensus on its definition [10]. Despite the need for a formalized definition, role is seen as a combination of a position or multiple positions, an individual who fills the position(s), as well as the responsibilities and behaviours of that individual. Role is meaningful only when the set of behaviours and responsibilities happen through communications. In other words, the domain of role is the set of interacting individuals and a role is revealed and developed via interactions among individuals. Similar to the kinetic energy K which is defined only for the moving objects, role is also defined only for interacting individuals. Thus, when there is no interaction or communication among people, role could not even be defined.

Consider two scenarios A and B in a classroom. In Scenario A, homework should be done individually and students are not allowed to collaborate on the homework solution. On the other hand, students have to build teams and work

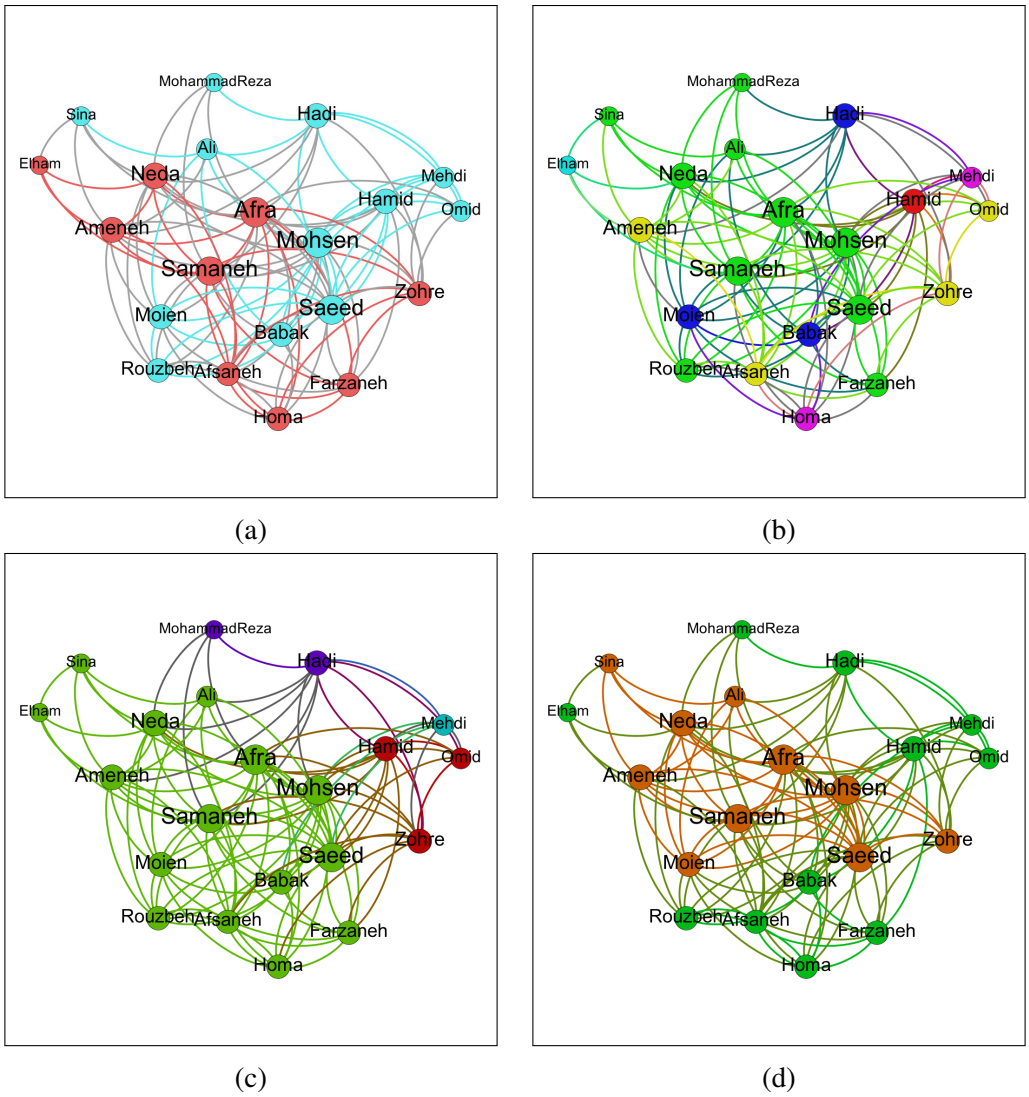


Figure 3.1: A group of lunchtime friends are shown in this figure where each person is represented by a node and edges between nodes indicate the "close friendship" relationships. Size and label of nodes represent their relative degree centrality score. Different communities of this network regarding various criteria are shown in this figure. (a) male and female communities, (b) hometown communities (Tehran, Isfahan, Mashhad, Tabriz, Hamedan, Ardebil, and Jahrom all cities of Iran), (c) major-based communities (CS,ECE,Civil,Chemical), and (d) degree-based communities (PhD,Masters).

on the homework together with their teammates in Scenario B. In the first scenario, there is no interaction among students and thus we cannot define or observe any role for them. However, in the second scenario, the students' reactions toward each other could be observed as their roles: some of them might participate more in the

discussions, while some others might prefer to talk less; some might not let others express their thoughts and want to push their own thoughts to be submitted; and some might have the ability to manage discussions and make other members of the team participate equally in the project. These students might have similar abilities and characteristics when doing their homework alone, but they show different behaviours when it comes to interacting with each other.

Having various definitions for the concept of role in sociology, here, we accommodate this concept into the network science by defining it in terms of the network terminology. A social network is modeled by a graph $G(V, E)$ where V is the set of vertices/nodes of the graph and E is the set of graph's edges. Here, the entities are associated with the nodes and the connections/interactions between the entities are represented with the edges. In addition, entities and connections between them can have multiple attributes depending on the network context. For instance, in the friendship network of classmates, students are vertices and their friendship relationships are the connections between them. Each student has several attributes like gender, age, hometown, etc. Connections between any two students also can have attributes like the beginning their friendship, the friendship strength, and the common interests in this relationship.

We categorize the information of a social network into *structural* and *non-structural* properties. Structural properties are related to the construction of the graph such as an entities' connections (edges), neighbourhood structure, and the entity's position in that structure. While, non-structural properties are other information not reflected in the construction of the graph like entities and connections' attributes and meta information of the graph. Considering all these properties, we define *role* in a network as follows:

Definition 1 *Role of an entity in a network is how it behaves toward others and its impact on others' attributes and structures. In a static network, role of an entity is determined by its structural and non-structural properties.*

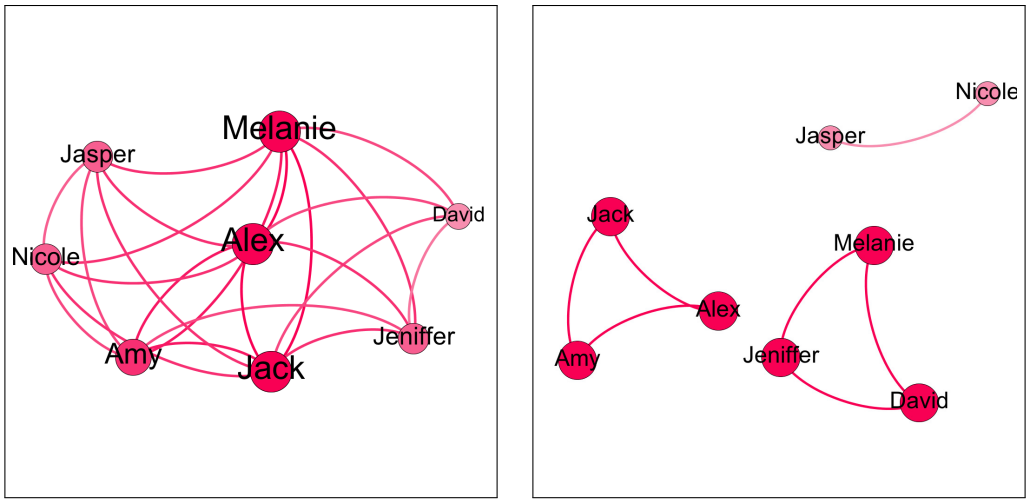
In temporal networks, structural and non-structural properties may change. Thus, we further can define *temporal-structural* and *temporal-non-structural* properties

to reflect the changes that happen in static features of the network. These temporal properties can also be used in identifying roles in that network. Since roles are important in the way they affect their environment, using temporal information from the network, which reflects the impact of entities, is useful in identifying roles. To this end, in identifying roles in a temporal network, temporal properties of entities can also be used together with other properties pointed out in Definition 1 to reflect the evolution and changes in the network through time.

3.3.2 Roles Defined within Our Framework

Human networks are intrinsically composed of multiple communities. In a social network with multiple communities, nodes' properties vary depending on whether the existence of communities is considered or neglected. From a social network analysis point of view, a node might be central in the whole network but not central in its community. Thus, we focus on studying the human based networks considering the existence of communities as their fundamental feature.

In social networks, communities can be either explicit or implicit. Explicit communities are built independently from their members and are based on a set of rules. In this case, people mostly join communities after the formation of the communities. Employees of a company or students participating in a course are examples of two explicit communities. Whereas the formation of implicit communities heavily depends on their members and connections. Thus, there is no external rule in building an implicit community. Implicit communities are built gradually as people come together. For example, friendship groups are implicit communities where there is no rule for individuals actions. In both cases of explicit and implicit communities, there should exist special individuals who manage and control the community. In the example of a classroom, it is the teacher or instructor who does such. For a company, managers are in charge and in the case of a friendship group, it is the person with more communication skills who brings together others and manages (implicitly) the relations to strengthen links between others leading to formation of a friendship group. These prominent individuals are even more highlighted when the size of a community is large. A group of 15 friends may not be able to hang out



(a) (b)

Figure 3.2: A group of 8 friends and their "close friendship" relationships are shown in this figure where size of nodes indicates their relative degree centrality score. The network in (a) shows the relations before Alex leaves the group, and (b) shows how the structure of the network destructs after Alex' leave.

if one is not responsible (either implicitly or explicitly) to plan for their gatherings. They may even decompose into smaller groups if the responsible person leaves or decides to become less active and no one else plays his/her role. Figure 3.2 is a real case happened to a friendship network where edges represent close friendships. Figure 3.2(b) is what happened to the community when Alex was not happy in this group anymore and started being inactive in the group. This split shows how important Alex was in keeping the group members together.

Following the above, we define roles for individuals in a social network considering their affiliations and positions in communities along with their interactions with other individuals. From the perspective of communities, in a network, individuals are of various types:

- with no affiliation to any community.
- connecting multiple communities.
- important members of a community.
- ordinary members of a community that form the majority

- non-important members of a community that do not noticeably affect the community

As a demonstration, Figure 3.3 shows several communities within a network as well as the position of individuals in each community and their importance in communities interactions. Based on these observations, here we define and consider four general fundamental roles, namely leader, outermost, mediator, and outsider as described in the following.

Definition 2 Leaders *are the outstanding individuals in terms of centrality or importance in each community. Leaders are commanders, directors, managers, rulers, controllers, pioneers, principals, presidents, authorities, administrators, or chairs of communities.*

Depending on how the community is constructed, the leader role might be stated explicitly or implicitly. For instance, in the example of the classroom, the instructor(s) and his/her assistant(s) are the explicit leading roles. In the friendship network among students of the same classroom, the student who is better known and more popular among all students implicitly has the leading role.

Definition 3 Outermost *are the small set of least significant individuals in each community whose influence and effect on the community are below the influence of the majority of the community members. Lexically, an outermost means further from the interior or center. Other words to express this class of individuals are peripheral, fringe, furthermost, remote, and borderline.*

In the example of the classroom, outermosts are those who are the least active students in the course. In the example of the friendship network, outermosts are those who barely communicate and hang out with others and have a few number of friends.

Definition 4 Mediators *are individuals who play an important role in connecting communities to each other in a network. They act as bridges between distinct communities. Mediators are negotiators, intermediaries, arbitrators, moderators,*

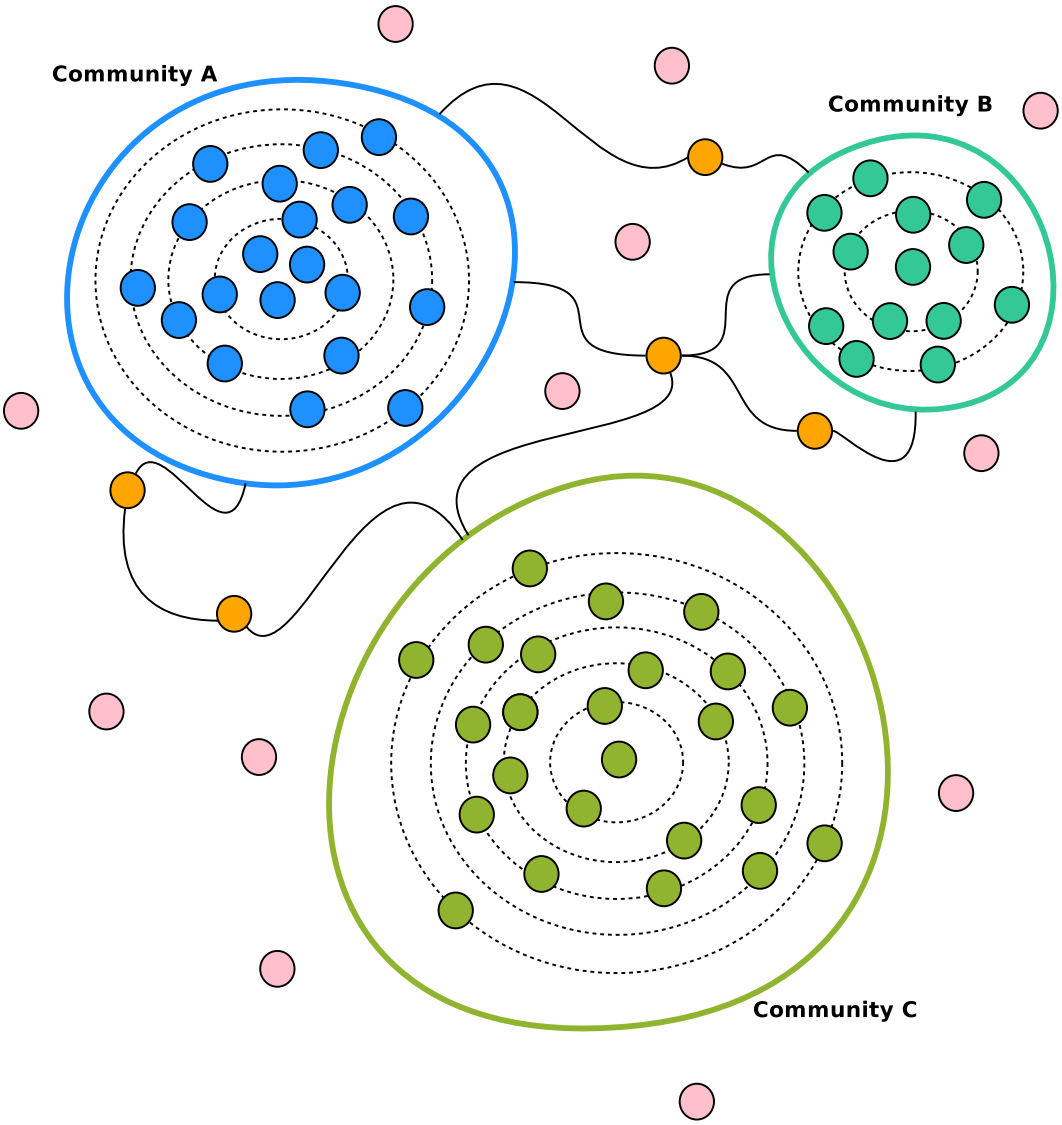


Figure 3.3: This figure shows the intuitive picture of how communities within network, and members within communities are placed. In this figure, 3 communities A, B, and C are shown. There are different kinds of nodes shown in this figure: some nodes are connecting communities to each other (orange nodes), some are with no connections or very week connections to communities (pink nodes), and other nodes are members of communities. Within each community, nodes are more central and more important as they are closer to the center. The more they are closer to the borders of communities, the more week and inactive they are.

*or hubs in a network. They might either be a member of a community (*inclusive mediator*) or not (*exclusive mediator*).*

Despite their affiliation to a specific community, mediators have many connections to multiple communities. Hence, they are controlling the communications

among various communities.

Some examples of mediators are diplomats in the relation between countries, publicity director of a company, a mixed French-Chinese student at a university could also be a mediator between the French students' association and the Chinese students' association. In all these examples, mediators are all inclusive (they belong to one of the communities), while there are occasions where mediators do not belong to any of the communities (they are exclusive). Back to the example of the classroom, when students build project teams to do their final project, the instructor(s) or teaching assistant(s) are mediators between project teams while they do not belong to a project team themselves.

Definition 5 Outsiders *are the individuals who are not affiliated with any of the communities in a network. They either have almost equal connections to different communities or have very few weak links to communities.*

Those outsiders who have very few connections are like outliers and the ones having many links to various communities at the same time are mediators (exclusive mediators). In addition to the examples of exclusive mediators, some examples of outliers are new migrants to a country, or new students registered in a course.

3.4 Structural Social Role Identification

Having a network with its communities explicitly known or extracted by a community mining algorithm, we propose methodologies for identifying the defined structural roles in the previous section.

Outsider

Having communities of a network, the most straightforward role to identify is the outsider. As each individual is either a member of a community or not, having the communities of a network known, individuals who do not belong to any community are outsiders.

Leader

Leaders in each community are the outstanding central members. To find the centrality of a node in its associated community, we choose an appropriate measure, for example one of the centrality measures described in Chapter 2. Then, we calculate this measure for all nodes inside a community giving us the centrality measure score for each node. Using the centrality measure scores, we are able to find the centrality measure probability distribution function (pdf) for each community. Analyzing the behaviour of the probability distribution function, we could identify leaders. For instance, in our experiments described in Chapter 4, nodes falling in the upper tail of the distribution function are identified as the community leaders. More details on our experiments can be found in Chapter 4.

Outermost

Similar to the leader, for identifying outermosts, we calculate how central are the individuals based on a measure, for example one of the centrality measures described in Chapter 2. Then, we compute a score for nodes of a community and use the pdf of the scores to identify outermosts. In our experiments described in Chapter 4, outermosts are identified as nodes falling in the lower tail of the pdf. Note that there is an important challenge for identifying outermosts using centrality measures. The intuition behind all centrality measures is to identify higher values, not lower ones. It means, high centrality scores for nodes infer their importance, however, low centrality scores does 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.

Mediator

Mediators are individuals who connect multiple communities and control communications between them. In comparison to mediators which are connectors between communities, nodes with high betweenness centrality are connectors between all nodes of the network. Using this similarity, we define two centrality measures based on betweenness centrality for extracting mediators in a network. These two vari-

ants of betweenness centrality are called *LBetweenness (LBC)* and *CBetweenness (CBC)*.

Prior to define LBetweenness, we need to define $LPath$ as follows:

*** LPath:** *LPath is the set of all shortest path between leaders of two distinct communities. More formally, LPaths is defined as*

$$LPath = \{l | startNode(l) \in leaderSet(c_i) \wedge endNode(l) \in leaderSet(c_j) \wedge c_i \neq c_j\} \tag{3.1}$$

where c_i and c_j are two arbitrary communities in the network.

*** LBetweenness:** *LBetweenness centrality for node v, denoted by $LBC(v)$, is the number of distinct LPaths that include v. If for each path $p \in LPath$, we define $I_l(p, v)$ to be 1 if v resides on p and 0 otherwise, then*

$$LB(v) = \sum_{p \in LPath} I_l(p, v). \tag{3.2}$$

*** CBetweenness:** *for each node v, its CBetweenness, called $CBC(v)$, counts the number of shortest paths between any two nodes of distinct communities that pass through node v. For each $p_{c_i,c_j} \in AllShortestPaths$ where two ends of p belong to communities c_i and $c_j \neq c_i$ respectively, we define $I_c(p_{c_i,c_j}, v)$ to be 1 if v resides on p_{c_i,c_j} and 0 otherwise. Thus:*

$$CBC(v) = \frac{1}{2} \sum_{c_i} \sum_{c_j \neq c_i} I_c(p_{c_i,c_j}, v) \qquad \textit{for undirected networks,} \tag{3.3}$$

$$CBC(v) = \sum_{c_i} \sum_{c_j \neq c_i} I_c(p_{c_i,c_j}, v) \qquad \textit{for directed networks.} \tag{3.4}$$

Figure 3.4 presents how LBC and CBC are different. In this figure a synthetic network composed of two communities is shown. Nodes l_1 and l_2 are respectively leaders of communities 1 and 2. As shown in Figure 3.4, leaders of two communities are connected to node A, while all other nodes in these two communities are connected to node$B$. Computing the LBC score, node A will get higher rank, however according to the CBC score, node B is more important.

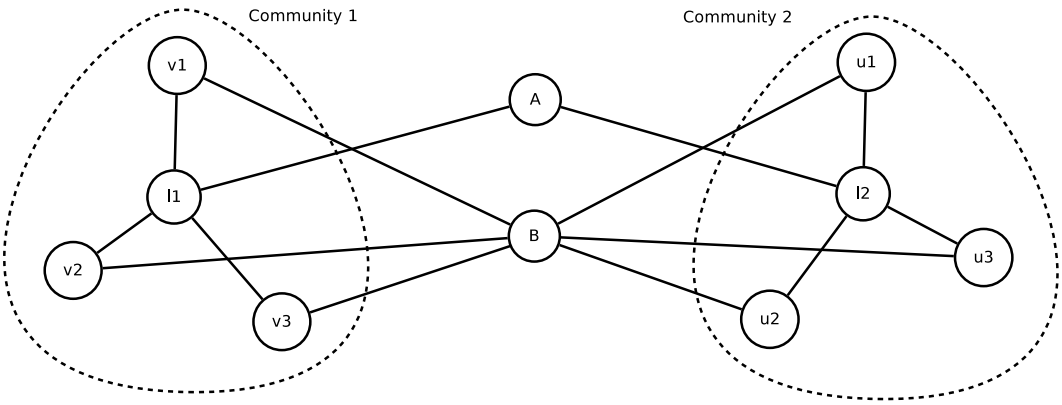


Figure 3.4: This figure presents a synthetic network consisting of two communities. As shown in the figure leaders of two communities (l_1 and l_2) are connected to the node A, while other nodes are all connected to node B. Computing LBC and CBC for all nodes of the graph, the results are as follows: $LBC(A) = 1$, $LBC(B, v_1, v_2, v_3, u_1, u_2, u_3) = 0$, $CBC(A) = 7$, $CBC(B) = 12$, $CBC(v_1, v_2, v_3, u_1, u_2, u_3) = 6$, and $CBC(l_1, l_2) = 3$.

In order to identify mediators, beside the two newly defined versions of betweenness centrality, it is also important that how many different communities are connected to each other through a node. For instance, assume a network consisting of 10 communities. Within this network, two nodes v_1 and v_2 have equal CBC value but different in how they are connected to communities. Node v_1 has connected three of the communities to each other, whereas node v_2 has connected all communities. Thus, in examining that how nodes are important in connecting communities, CBC values alone is not sufficient. To this end, we incorporate the number of connected communities through a node in our role mining framework. Hence, we define another ranking score called *diversity score*. Diversity score counts the variants of communities a node is connected to.

*** Diversity Score:** *This score shows how many different communities are connected through a specific node v. Depending on the definition, we calculate the diversity score in two different ways:*

- *Diversity score of node v, $DS_{count}(v)$, is defined as the number of communities connected to each other through node v. For each node v, we first define $I_d(c_i, v)$ to be 1 if $\exists\, c_j, a, b, p : a \in c_i \wedge b \in c_j (\neq c_i) \wedge p \in SPath(a, b) \wedge v \in p$ and 0 otherwise. Here, c_i and c_j represent two communities and $SPath(a, b)$*

is the set of shortest paths between nodes a and b. Using $I_d(c_i, v)$, $DS(v)$ is then defined as follows:

$$DS_{count}(v) = \frac{1}{2}\sum_{c_i} I_d(c_i, v) \qquad \textit{for undirected networks,} \tag{3.5}$$

$$DS_{count}(v) = \sum_{c_i} I_d(c_i, v) \qquad \textit{for directed networks.} \tag{3.6}$$

- *Diversity score can also be defined to count pairs of communities having 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 $\exists\, a, b, p : a \in c_i \wedge b \in c_j (\neq c_i) \wedge p \in SPath(a, b) \wedge v \in p$. Now,*

$$DS_{pair}(v) = \frac{1}{2}\sum_{c_i}\sum_{c_j \neq c_i} I_d(c_i, c_j, v) \qquad \textit{for undirected networks,} \tag{3.7}$$

$$DS_{pair}(v) = \sum_{c_i}\sum_{c_j \neq c_i} I_d(c_i, c_j, v) \qquad \textit{for directed networks.} \tag{3.8}$$

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

Using a combination of the above metrics, the mediacy of nodes can be computed as a function of $f(LB, CB, DS_{count}, DS_{count})$. We rank nodes by their mediacy score. Then, Algorithm 1 is used to find highly ranked nodes connecting the maximum number of communities to each other.

3.5 Summary

In this chapter, we proposed the SSRM framework to study roles of individuals in a social network. The framework is built upon the assumption of existence of non-overlapping communities forming the network. Using the structure of communities, we defined four fundamental roles namely leader, outermost, mediator, and outsider. Furthermore, we propose methodologies to identify these roles. Based on the definitions, outsiders are straightforwardly identified using the information

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
```

about community memberships. The remaining three roles are identified using appropriate measures are necessary for ranking nodes in the communities (to identify leaders and outermosts) and the whole network (to identify mediator). We chose a measure for computing scores for nodes in each community. Based on that measure, leaders and outermosts are identified considering the distributions of nodes' score computed by the chosen measure. In the identification of mediators, we chose a network-wide measure to compute score of nodes. Similarly mediators are identified considering the distribution of the scores or the Algorithm 1.

Chapter 4

Case Study: The Enron Email Dataset

Those that much covet are with gain so fond,
For what they have not, that which they possess
They scatter and unloose it from their bond,
And so, by hoping more, they have but less;
Or, gaining more, the profit of excess
Is but to surfeit, and such griefs sustain,
That they prove bankrupt in this poor-rich gain.

William Shakespeare, The Rape of Lucrece

4.1 What is Enron?

Enron was an American energy company formed in 1985 when InterNorth bought Houston Natural Gas (HNG) and merged these two companies together. Kenneth Lay, the former HNG CEO, became Enron CEO after Samuel Segnar, the ex-InterNorth and first CEO of Enron, departed 6 months after the merge of the two companies [77].

Enron constructed the first nationwide gas pipeline in the United States.

Figure 4.1: Enron Logo

Beside their gas-related business, the company expanded its focus to many other products including petrochemicals, plastics, power, pulp and paper, steel, weather risk management, oil and liquid natural gas transportation, broadband, principal investments, risk management for commodities, shipping and freight, streaming media, and also water and wastewater [77]. Enron was one of the world's largest gas companies with more than 22000 employees in over 40 countries [26, 48].The Fortune 500 magazine ranked Enron as "America's most innovative company" for 6 consecutive years. In 2000, they claimed about $101 billion revenue during that year and the company's stock price reached to a maximum of $99 US. These tremendous successes were made through misleading accounts to hide debts and losses of the company. In October 2001, they announced $638 million losses in the third quarter and $1.2 billion reduction in shareholders equity. Consequently, the company filed for bankruptcy on the December 2^{nd}, 2001.

4.2 Network Characteristics

The Enron corpus of emails from about 150 employees was made public by the U.S. Federal Regulatory Commission (FERC) during the legal investigation of Enron in May 2002. The corpus includes about 250,000 emails from 1998 to 2002 [65]. It is the only real email dataset available for research and has been used in the fields of data mining, social network analysis, text mining, natural language processing, and organizational studies [79, 66, 47, 20, 40, 8].

In this work, we use the Enron communication network built upon the email exchanges in each month of the year 2001 with a total of 285 nodes and 23559 edges [68, 69]. More specifically, the network is made up of 12 timeframes each representing a month from January 2001 to December 2001. In each timeframe, a node represents an Enron employee who has exchanged emails in that month, and an edge between two individuals is an indicator of at least one email communication between them in any direction in that timeframe. Thus, the resulting network is an undirected, unweighted communication network.

4.3 Experiments

4.3.1 Experimental Setups

The structural social role mining framework is built upon the assumption of existence of communities. Due to this assumption, we need to find communities prior to applying our structural social role mining framework. For this purpose, we use *Local Community Mining Algorithm* [17] to find communities. Having the communities extracted, we can apply the structural social role mining framework in order to identify roles.

Subsequently, we study temporal changes of roles in the Enron communication network. Since the Enron communication network is made up of multiple timeframes, we are able to study temporal events in this network. To this end, we use the community events introduced and identified by Takaffoli et al. in [67, 68] on the same dataset as an example of temporal events. Furthermore, we observe the mutual effects of changes of the roles we define and the community events defined by Takaffoli et al. on the Enron communication dataset.

The code written for identifying roles is in Java using Jung library[1]. In addition, Gephi[2] and Python's matplotlib library[3] are used for visualizing networks and drawing plots.

4.3.2 Choosing a centrality measure for identifying roles

Centrality Measure: Leader and outermost

Degree and closeness centralities are good candidates for ranking individuals in a community in order to identify leaders and outermosts. Figure B.1 to Figure B.12 in Appendix B, show degree distributions of nodes in communities in each timeframe for the year 2001. Distributions are plotted for communities with more than 30 nodes in each timeframe. In addition, normal distributions with the same mean and variance values are plotted for each community in the same figure. As shown in Figure 4.2, degree distributions are more like power-law distribution than normal

[1] http://jung.sourceforge.net/
[2] https://gephi.org/
[3] http://matplotlib.org/

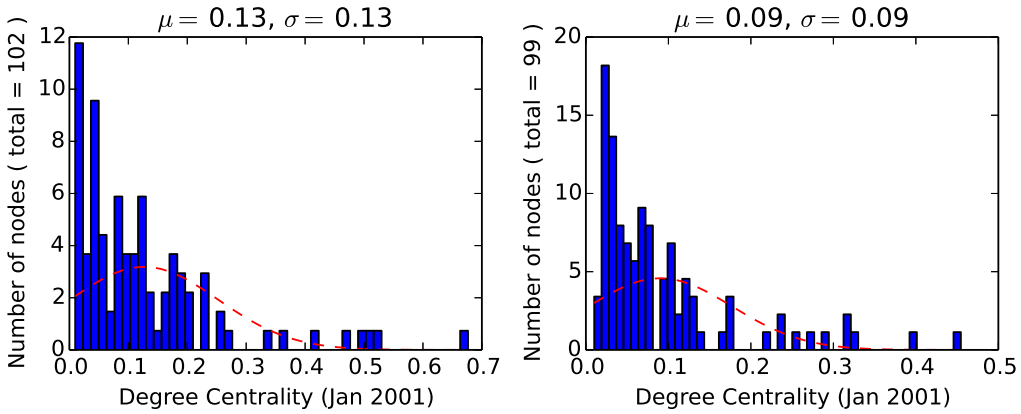


Figure 4.2: Plot of Communities' Degree Distribution for January 2001

distribution as they have mostly one tail (the upper tail). The shape of degree distributions interprets the fact that the majority of the population in each community has low degrees. Although, the long upper tail of degree distributions can be used to identify leaders, outermosts cannot be simply identified using degree distributions of communities due to the large number of low degree nodes.

To find a measure that can be used to identify both leader and outermost roles, we consider closeness centrality rather than degree centrality. In spite of the power-law behaviour of degree centrality distributions, closeness centrality distribution of individuals within a community follows a normal-like distribution as shown in Figure 4.3. Since closeness distribution has two tails in each community of our dataset, it can be effectively used to identify both leaders and outermosts. Distributions of degree and closeness centrality for all timeframes from January 2001 to December 2001 are provided in Appendix B.

Centrality Measure: Mediator

We use measures defined in Chapter 3 to compute the degree of mediacy of a node. Since the number of leaders and the number of shortest paths between them are small in our dataset, we observe that LBC is not an appropriate indicator for mediators. Hence, we use CBC to identify mediators in our experiments. Although CBC is more expensive in terms of time complexity, it identifies mediators that could not be identified by LBC in our dataset. LBC only considers shortest paths

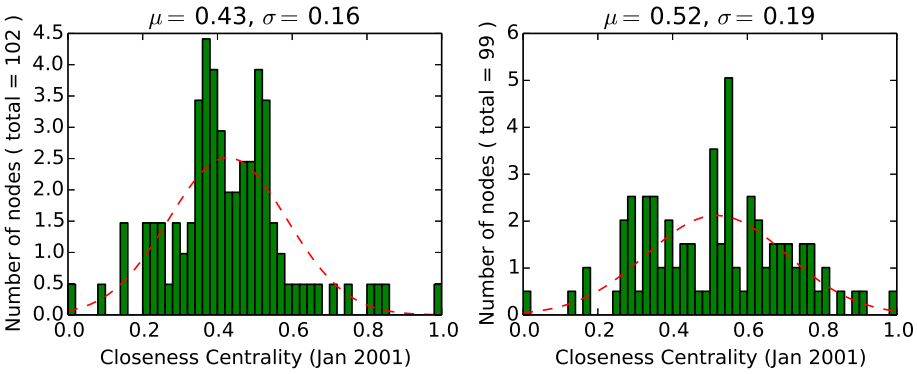


Figure 4.3: Plot of Communities' Closeness Distribution for January 2001

between community leaders, whereas CBC takes into account all shortest paths between community members. As a result, mediators found by CBC more strongly connect members of two communities, while mediators found by LBC connect community leader to eachother. Thus, LBC can be a better and more efficient choice for identifying mediators in situations where members of communities other than the leaders do not have many connections to individuals outside their community. In addition to time complexity, CBC has another challenge to overcome. That is, 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 leads to more shortest paths between them.

Figure 4.4 shows a network consisting of four communities, two larger communities and two smaller ones. The larger communities 1 and 2 connect through node R and the communities 3 and 4 connect through node S. Connections between nodes are not shown in Figure 4.4 and edges in the figure represent shortest paths between nodes passing through nodes R and S. In addition, edges in the network are coloured in *red* or *blue*. When a node (such as node A) in community 1 is connected to another node (such as node G) in community 2 through node R and edges of the same colour, it means that there exists a shortest path between those two nodes (nodes A and G). Thus, paths and colours in Figure 4.4 are interpreted as shortest path passing through node R (S) between nodes of different communi-

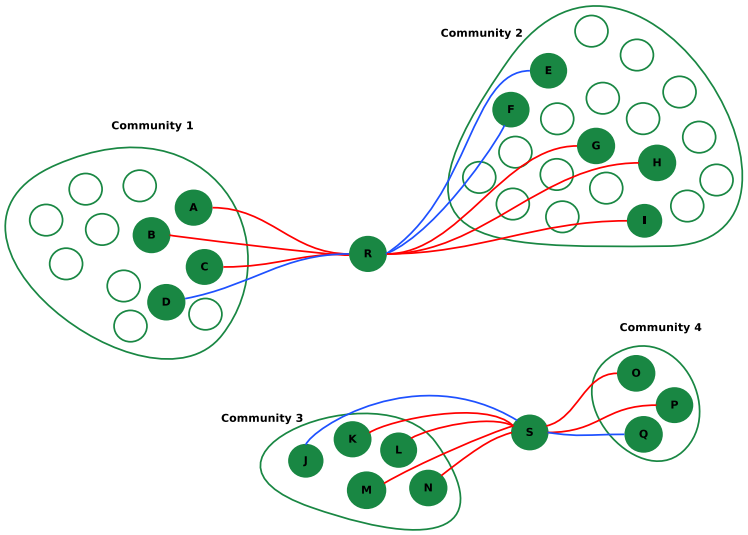


Figure 4.4: In this figure, node R, connecting two large communities 1 and 2 is compared to node S connecting two smaller communities 3 and 4. Connections between nodes are not shown and edges of the graph represent shortest paths between their members passing through nodes R and S.

ties. For instance, there are shortest paths between pairs of nodes (G, A), (G, B), (G, C), (H, A), (H, B), (H, C), (I, A), (I, B), (I, C), (E, D), (F, D) passing through node R. Same scenario happens for communities 3 and 4. Altogether, node S lies on 8 shortest path between communities 3 and 4, whereas node R lies on 11 shortest paths between communities 1 and 2. As a result, in the CBC ranking node R comes first, while node S is more strongly connecting communities 3 and 4 to each other. More precisely, node S lies on 8 shortest paths from 15 (5×3) possible shortest paths between communities 3 and 4, while node R lies on 11 shortest paths from 228 (12×19) possible shortest paths between communities 1 and 2. To compensate for this effect, we use *normalized* CBC as follows:

$$NCBC(v) = \frac{1}{2} \sum_{c_i} \sum_{c_j \neq c_i} \frac{I_c(p_{c_i,c_j}, v)}{min(size(c_j), size(c_j))} \qquad \text{for undirected networks,} \tag{4.1}$$

$$NCBC(v) = \sum_{c_i} \sum_{c_j \neq c_i} \frac{I_c(p_{c_i,c_j}, v)}{min(size(c_j), size(c_j))} \qquad \text{for directed networks.} \tag{4.2}$$

Figure 4.5 compares how nodes' ranking may differ using $NCBC$. As shown

in Figure 4.5, two nodes connecting the small community (in the lower left part of the network) to the rest of the graph will get higher ranks using the $NCBC$.

While two mediators may have equal scores, they can be substantially different in terms of number of distinct communities they connect to each other. Assume a network consisting of 10 communities of the same size and two mediators M_1 and M_2, M_1 lies on 100 shortest paths connecting two communities to each other. Equally, M_2 lies on 100 shortest paths but connecting all 10 communities to each other. Although M_1 and M_2 are similar in the number of shortest paths between distinct communities passing through them, node M_2 connects communities more globally than M_1. Thus, these two nodes cannot be evaluated equally.

Figure 4.6 presents distributions of NCB and $NCB \times DS_{count}$ for January 2001. Complete distribution plots for all timeframes are presented in Figure B.25 to Figure B.36 in Appendix B. As depicted in the figures, considering diversity score does not change the trend of the diagram, and most importantly the shape of its tail in the Enron communication network. However, a noticeable change happens in the middle part of the histogram[4]. To this end, we define *mediator score* as the multiplication of the node's $NCBC$ and diversity score as follows:

$$MS(v) = NCB(v) \times DS_{count}(v). \tag{4.3}$$

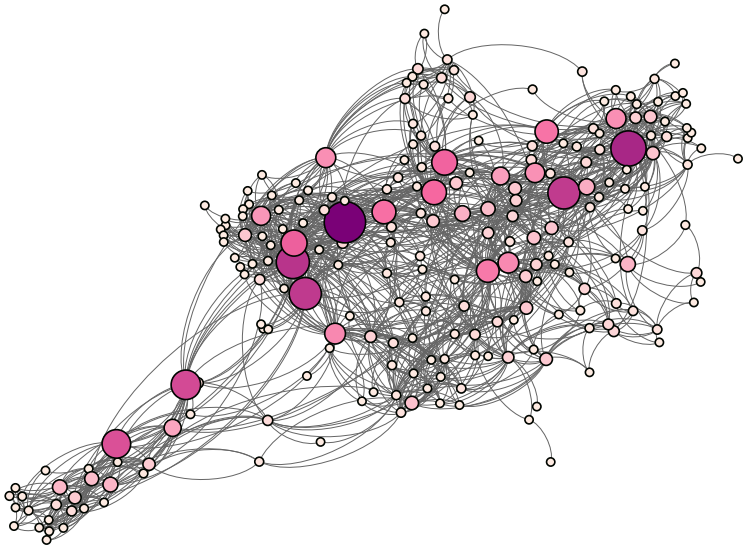
4.3.3 Identifying roles

After choosing a metric to rank nodes, i.e. closeness centrality for outermost and leader and mediator score for mediators, we develop the following methods to extract roles from the ranked lists.

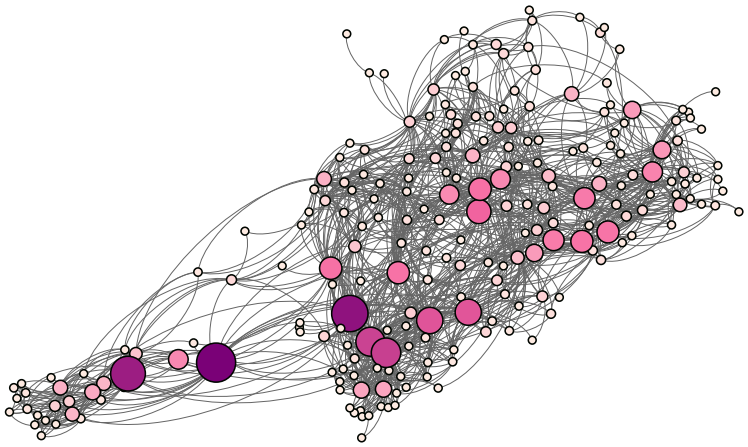
Identification: Leaders and Outermosts

The closeness distribution used for identifying leaders and outermosts in the Enron communication network is close to the normal distribution. Since in a normal distribution almost 95% of the population lies in the interval $[\mu - 2\sigma, \mu + 2\sigma]$,we respectively use $\mu + 2\sigma$ as the upper and $\mu - 2\sigma$ as the lower thresholds to identify

[4]Note that depending on the dataset, the effect of diversity score can be more significant.



(a) Nodes of the Enron communication network where size of the nodes represents their CBC before normalization.



(b) Nodes of the Enron communication network where size of the nodes represents their CBC after normalization.

Figure 4.5: Relative CBC scores for a set of network nodes are shown in this figure. Size and colour of nodes indicate their rank in the (normalized) CBC scoring list. More specifically, larger size and darker colour mean higher rank in the list. (a) shows graphical ranking of the nodes by their CBC scores, while (b) depicts the ranking by $NCBC$ scores. As shown in the figure, the ranking of nodes that connect the small community in the lower left part of the network improves when $NCBC$ is used (i.e. they become larger in size and darker in color).

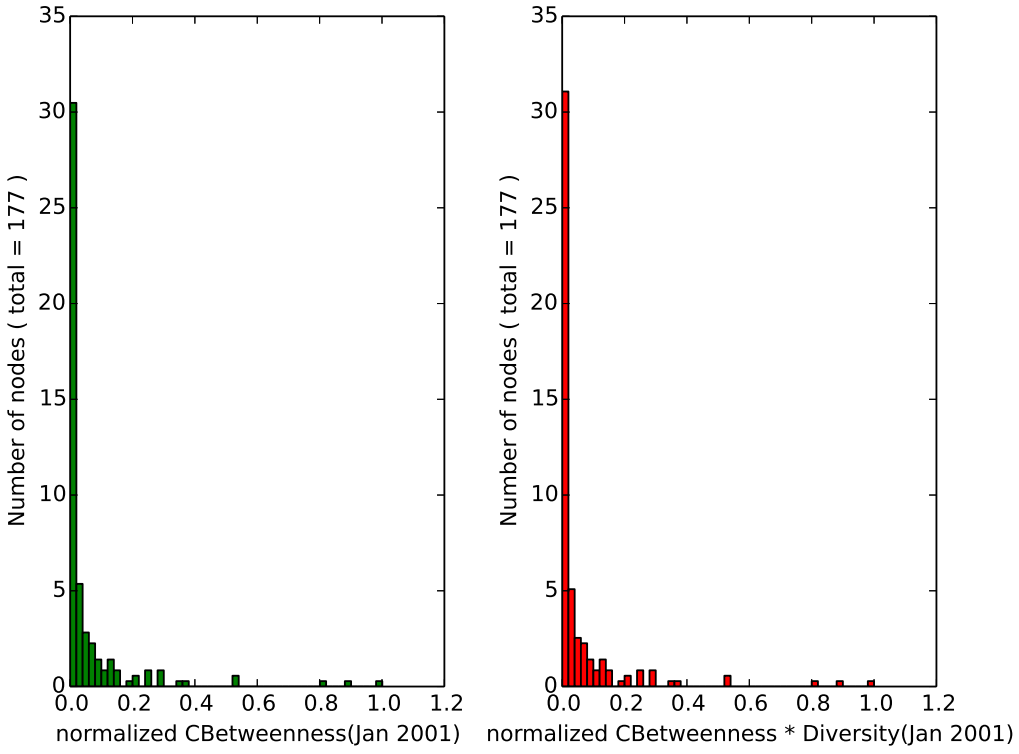


Figure 4.6: Distribution of NCB and $NCB \times DS_{count}$ scores for January 2001

leaders and outermosts. Figure 4.7 shows communities of the Enron communication network in one timeframe (August 2001) and visually presents how central a node is within its community by its size. As mentioned, centrality of nodes are computed based on the closeness centrality measure. Using the nodes' centrality scores, we form the closeness distribution and find the leaders and outermosts according to aforementioned thresholds.

Leaders identified in August 2001 are: Kimberly Watson, Rhonda L. Denton, Jannette Elbertson, Becky Spencer, Billy Lemmons, Kenneth Lay, Susan J. Mara, Jeff Dasovich, Richard Shapiro, and Ginger Dernehl. Being a leader in the network shown in Figure 4.8 translates into having a high average of short email distance to other individuals in the community. Among these names, Kenneth Lay was founder, CEO, chairman, and chief executive officer of Enron. Kimberly Watson was one of the directors and an influential person in the company based on the information we found about her in two meeting minutes. "Kim Watson made a motion to approve the expense of a special grant of Enron stock options as detailed by Robert Jones

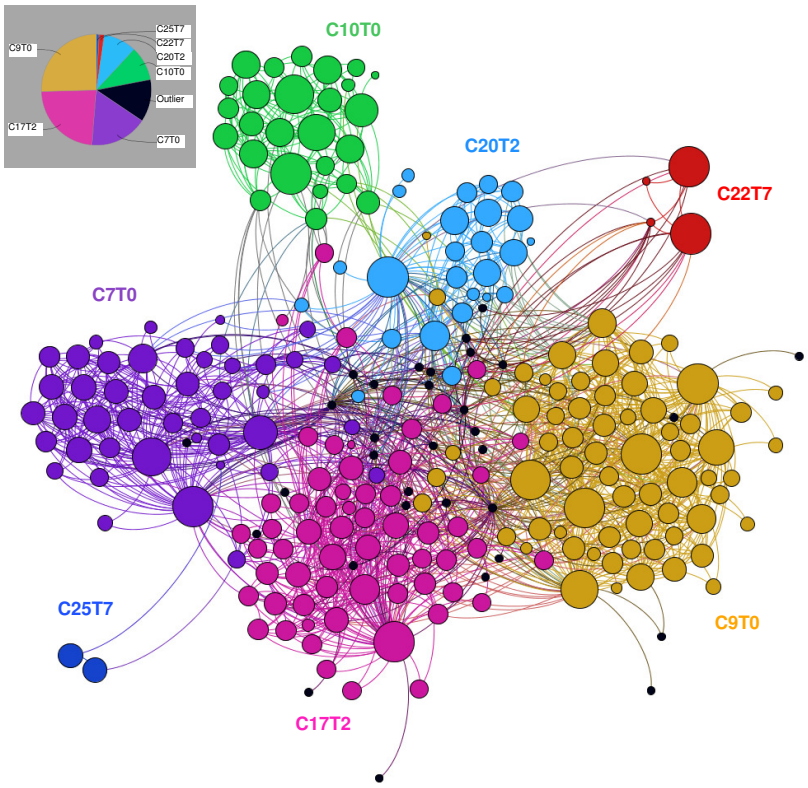


Figure 4.7: The communities within the Enron communication network in August 2001. Colours represent communities except for black that represents the outliers. Moreover, size of the nodes indicates how central (based on closeness centrality) a node is within its community. The bigger the size of the node, the more central it is in the community.

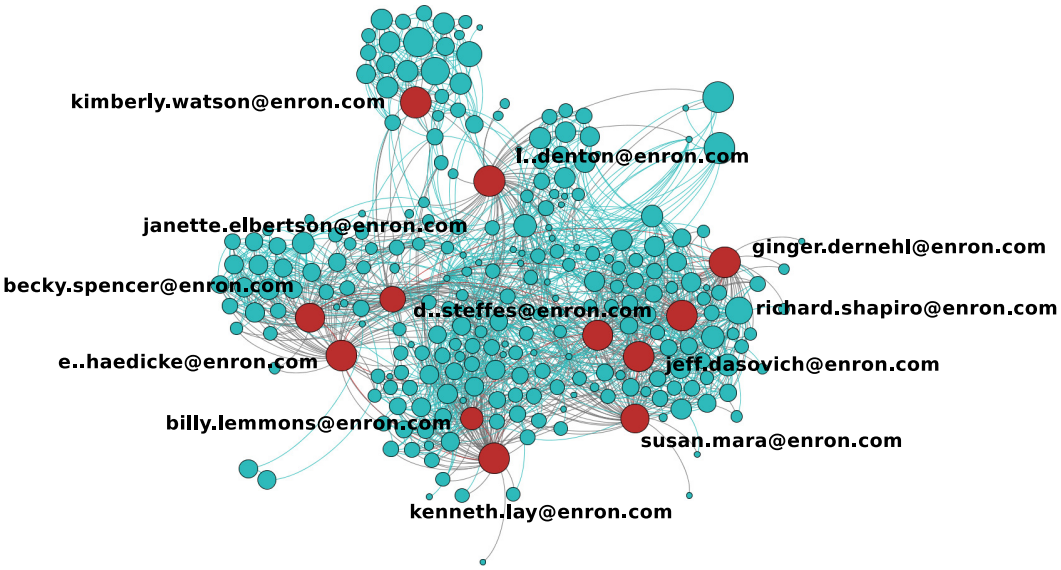


Figure 4.8: Leaders of communities are shown as red nodes in the Enron communication network during the month of August 2001. Similar to Figure 4.7, size of the nodes shows their centrality in their communities.

to Gary Hugo, Debbie Weir and Stephanie Olszenski. The motion carried by board vote." from ENRON FEDERAL CREDIT UNION MINUTES OF A REGULAR MEETING in 21^{st} of August 2000 [63]. In another meeting minute, it is mentioned that Ms. Kimberly Watson made a motion to approve a Treasurer's Report presented by Ms. Kennedy which was carried by Board vote in 17^{th} of July 2000 [64]. Information about Becky Spencer and Ginger Dernehl has not been found, but other identified leaders are important people as well.

Figure 4.9 depicts the outermosts in August 2001. As shown in this figure, size of the nodes having the outermost role is very small referring to their low centrality score within their communities. In the Enron communication dataset, outermosts are less frequent than leaders according to the closeness centrality distributions in Figure B.13 to Figure B.24. If there is no node in a community with smaller centrality score than $\mu - 2\sigma$, we would not identify any outermost for that community. In Figure 4.9, only two of the communities have outermost roles. This may also happen for leading role when there is no node with a centrality score greater than $\mu + 2\sigma$.

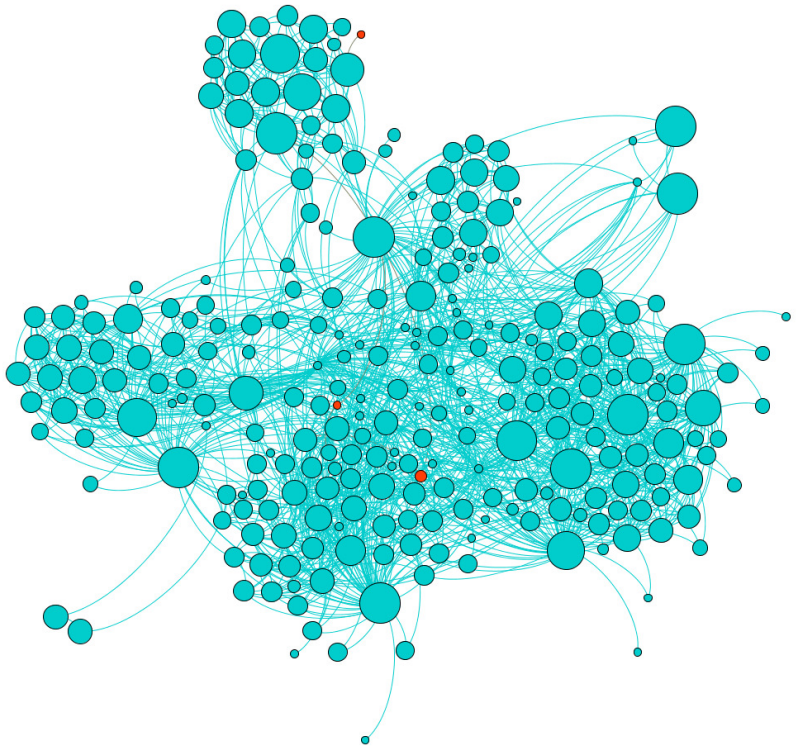


Figure 4.9: Enron communication network with identified outermost roles within communities. Red nodes are indicator of outermost roles. As it is shown in this figure, only two communities have outermost role. The notion of the role outermost by definition refers to the individuals whose importance is reasonably below the importance of the majority of nodes in that community. Thus, when the majority of nodes have low scores according to their importance, no node will be identified as an outermost which is the case for most of the communities in this network. From top to bottom, outermosts are Ava Garcia (probably an assistant according to the body of some emails), Leslie Reeves a module manager, and Shirley Crenshaw (probably an assistant).

Identification: Mediators

Figure 4.10(a) shows communities of the Enron communication network in October 2001. The colour of nodes indicates their community, while the size of nodes shows their relative mediator score. Having the ranked list of nodes by their mediator score, we need a method to extract mediators. In this experiment, we use two methods to this end. In the first method, we use the MedExtractor (Algorithm 1) to identify mediators. Mediators found by this algorithm are high ranked individuals according to their mediator score that all together connect the communities in the network. These nodes are shown in red in Figure 4.10(b).

In the second method, we use the distribution of the mediator score to identify mediators. As shown in Section 4.3.2, the mediator score distributions are close to a power-law distribution. A power-law distribution is followed by the Pareto principle or the 80-20 rule. The 80-20 rule states that 20% of the causes result in 80% of the effects. Which means that the tail of the distribution posses a large portion of the values. In the case of Enron communication network, it means that 20% of the individuals in the network mediate 80% of the communications between communities. Based on this fact, we use the tail of the mediator score distribution to identify mediators. However, tail of the mediator score distributions for the Enron communication network is very sparse. The reason for sparse tails could be originated from 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. Figure 4.10(c) depicts the set of 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 most of the shortest paths between communities pass through them.

MedExtractor (Algorithm 1) determines the minimum number of mediators that connect all communities to each other, while using the distributions, we find nodes controlling most of the connections between communities. Thus, on the condition that it is important to minimize the size of mediators that at the same time connect the maximum number of communities, MedExtractor (Algorithm 1) gives better

results. However, if the goal is to target all nodes that are controlling most of the shortest paths between communities, using the 80-20 rule on the distribution meets our goal. As shown in Figure 4.10(b) and 4.10(c), nodes that are selected by the MedExtractor (Algorithm 1) are a subset of nodes identified by the 80-20 rule. Thus, depending on what we expect from mediators, one of the aforementioned methods can be used.

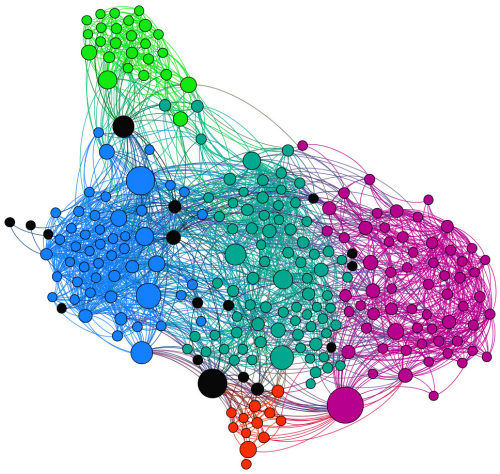
4.3.4 Roles Changes

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 leader or mediator at least once in different timeframes.

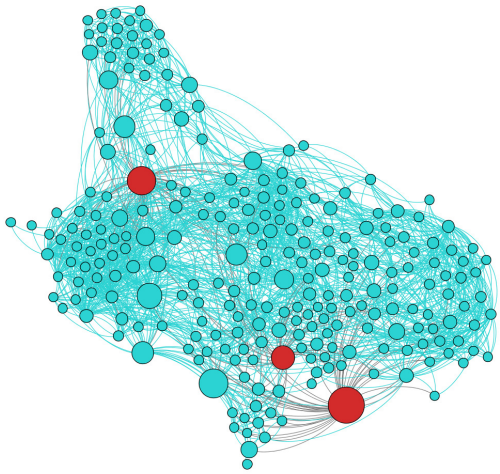
We present change of roles by a table where rows and columns are respectively nodes and timeframes. The intersection of rows and columns build cells. Each cell has information about the role of the associated node in the respective timeframe. For better visualization, cells are coloured to represent nodes' relative strength in having a special role.

Figure 4.11 is the associated table with the leading role. This figure indicates how leader nodes change their leadership through time. We track nodes' roles after the first time they become leaders. Thus, whether they are present in the network or not before their first leading role is not depicted in Figure 4.11. Information presented in Figure 4.11 indicates that some nodes serve as leaders in most of the timeframes. Some other nodes are present in the network during few timeframes and they act as leaders in those timeframes. On the other hand, some nodes are constantly leaders in early timeframes while some others are constantly leaders in later timeframes. Intuitively, being a constant leader over more timeframes could mean that a node is more important and affective in its community.

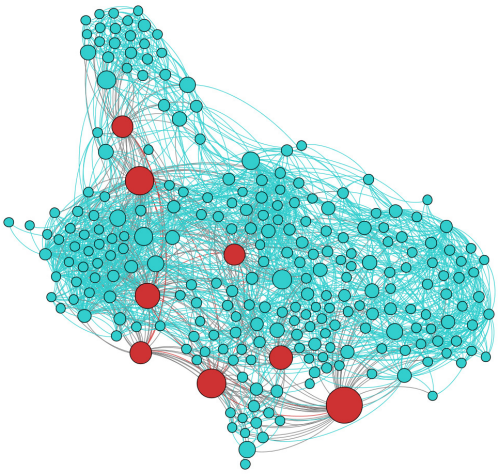
Similar to the time tracking figure for the leading role, Figure 4.12 presents how mediators change their role through time. Unlike the leading role where several nodes attain their leadership over several timeframes, we observe more fluctuations over the mediating role over time. According to Figure 4.12, mediators identified



(a) Communities within October 2001 timeframe. Size of the nodes depicts their relative mediator score



(b) Mediators (red nodes) found by MedExtractor (Algorithm 1)



(c) Mediators (red nodes) found using the mediator score distribution

Figure 4.10: (a) Relative mediator scores, community membership, mediator sets.

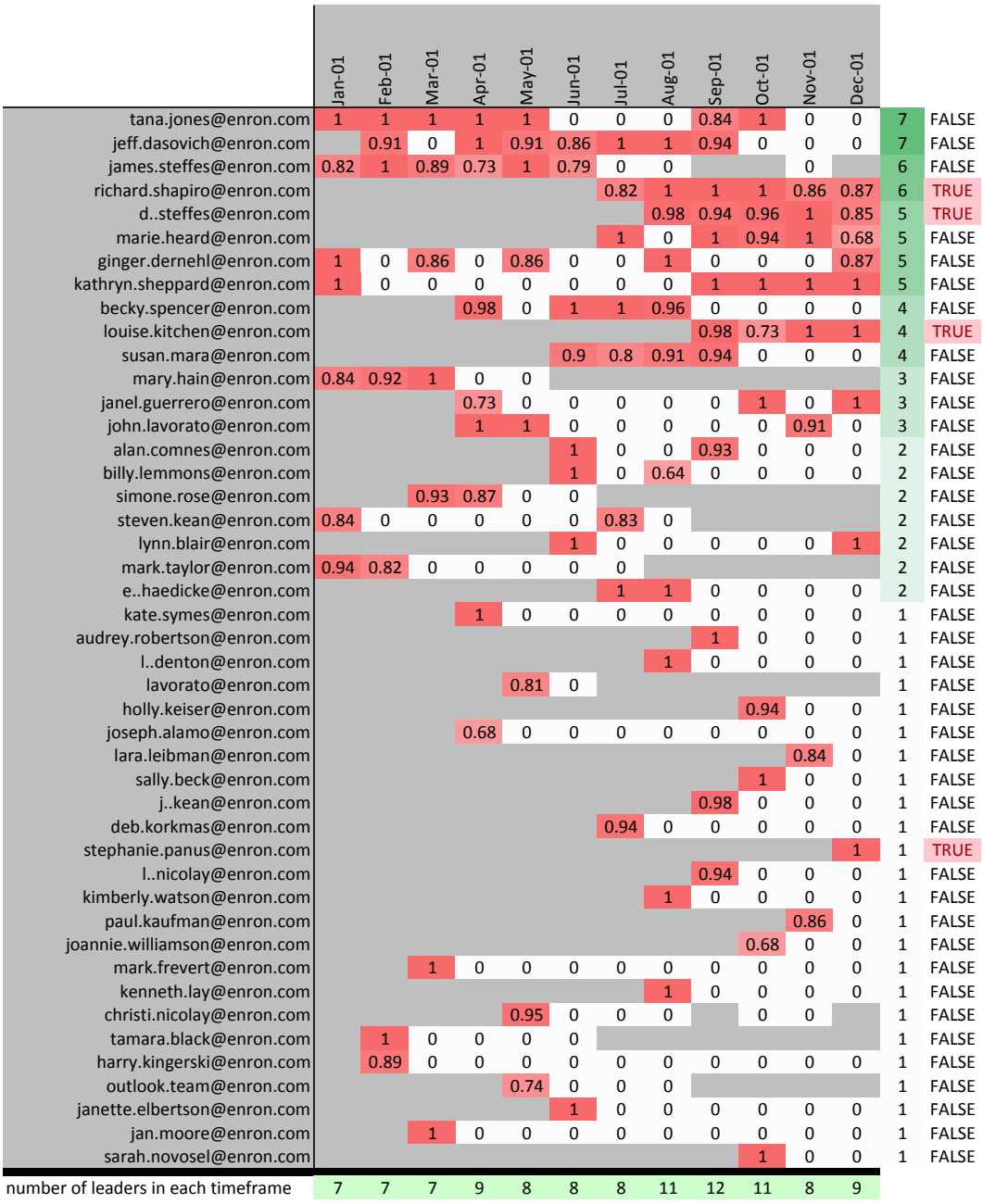
for the months July, August, and September express higher mediator scores. This means that more conversations have happened between different communities of Enron employees in these three months. Since we start tracking nodes after the first time they become a mediator, Figure 4.12 does not contain information about timeframes prior to acquiring the mediating role.

Overall, Figure 4.13 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 act as mediator and leader. This figure gives us knowledge regarding the role transition of each node. For example, Jeff Dasovich was a mediator in January 2001, then transforms into simultaneous leader and mediator. Later on, he becomes an outlier and then regains his leading and mediacy. One more example is Kimberly Watson who is an outermost in January, later on becomes an outlier and in 3 timeframes becomes a leader.

Tracking such temporal changes results in interesting information that helps in analyzing phenomena happening in the network. For instance, Steve J. Kean who served as Kenneth Lay's chief of staff in Enron's office of the chairman is an outsider in August 2001 and becomes a leader in September 2001 after Lay becomes the CEO again. Based on the role transitions, we can define a set of dynamic roles such as an outlier turning into a mediator or an outermost turning into a leader. We do not come up with different names for multiple combinations of role transitions. However, we use these transitions (dynamic roles) to analyze events happening for communities in the next section.

4.3.5 Role Transitions and Community Events

As seen in the previous section, the role of a node may change through time. Simultaneously, structures of communities may also change through time. For example, a new community may form, a former community may split or dissolve, or multiple former communities may merge into a larger one. We name role changes as *role events* and community changes as *community events*. Community events are



	Jan-01	Feb-01	Mar-01	Apr-01	May-01	Jun-01	Jul-01	Aug-01	Sep-01	Oct-01	Nov-01	Dec-01		
tana.jones@enron.com	1	1	1	1	1	0	0	0	0.84	1	0	0	7	FALSE
jeff.dasovich@enron.com		0.91	0	1	0.91	0.86	1	1	0.94	0	0	0	7	FALSE
james.steffes@enron.com	0.82	1	0.89	0.73	1	0.79	0	0			0		6	FALSE
richard.shapiro@enron.com							0.82	1	1	1	0.86	0.87	6	TRUE
d..steffes@enron.com								0.98	0.94	0.96	1	0.85	5	TRUE
marie.heard@enron.com							1	0	1	0.94	1	0.68	5	FALSE
ginger.dernehl@enron.com	1	0	0.86	0	0.86	0	0	1	0	0	0	0.87	5	FALSE
kathryn.sheppard@enron.com	1	0	0	0	0	0	0	0	1	1	1	1	5	FALSE
becky.spencer@enron.com				0.98	0	1	1	0.96	0	0	0	0	4	FALSE
louise.kitchen@enron.com									0.98	0.73	1	1	4	TRUE
susan.mara@enron.com						0.9	0.8	0.91	0.94	0	0	0	4	FALSE
mary.hain@enron.com	0.84	0.92	1	0	0								3	FALSE
janel.guerrero@enron.com				0.73	0	0	0	0	0	1	0	1	3	FALSE
john.lavorato@enron.com				1	1	0	0	0	0	0	0.91	0	3	FALSE
alan.comnes@enron.com						1	0	0	0.93	0	0	0	2	FALSE
billy.lemmons@enron.com						1	0	0.64	0	0	0	0	2	FALSE
simone.rose@enron.com			0.93	0.87	0	0							2	FALSE
steven.kean@enron.com	0.84	0	0	0	0	0	0.83	0					2	FALSE
lynn.blair@enron.com						1	0	0	0	0	0	1	2	FALSE
mark.taylor@enron.com	0.94	0.82	0	0	0	0	0						2	FALSE
e..haedicke@enron.com							1	1	0	0	0	0	2	FALSE
kate.symes@enron.com				1	0	0	0	0	0	0	0	0	1	FALSE
audrey.robertson@enron.com									1	0	0	0	1	FALSE
l..denton@enron.com								1	0	0	0	0	1	FALSE
lavorato@enron.com					0.81	0							1	FALSE
holly.keiser@enron.com										0.94	0	0	1	FALSE
joseph.alamo@enron.com				0.68	0	0	0	0	0	0	0	0	1	FALSE
lara.leibman@enron.com											0.84	0	1	FALSE
sally.beck@enron.com										1	0	0	1	FALSE
j..kean@enron.com									0.98	0	0	0	1	FALSE
deb.korkmas@enron.com							0.94	0	0	0	0	0	1	FALSE
stephanie.panus@enron.com												1	1	TRUE
l..nicolay@enron.com									0.94	0	0	0	1	FALSE
kimberly.watson@enron.com								1	0	0	0	0	1	FALSE
paul.kaufman@enron.com											0.86	0	1	FALSE
joannie.williamson@enron.com										0.68	0	0	1	FALSE
mark.frevert@enron.com			1	0	0	0	0	0	0	0	0	0	1	FALSE
kenneth.lay@enron.com								1	0	0	0	0	1	FALSE
christi.nicolay@enron.com					0.95	0	0	0		0	0		1	FALSE
tamara.black@enron.com		1	0	0	0	0							1	FALSE
harry.kingerski@enron.com		0.89	0	0	0	0	0	0	0	0	0	0	1	FALSE
outlook.team@enron.com					0.74	0	0	0					1	FALSE
janette.elbertson@enron.com						1	0	0	0	0	0	0	1	FALSE
jan.moore@enron.com			1	0	0	0	0	0	0	0	0	0	1	FALSE
sarah.novosel@enron.com										1	0	0	1	FALSE
number of leaders in each timeframe	7	7	7	9	8	8	8	11	12	11	8	9		

Figure 4.11: This figure shows how leader nodes change their leadership through time. For each node in each timeframe, there is either a 0, a number greater than 0, or a grey cell. 0 means that the node is present in that timeframe but not as a leader, the number means the node is leader and shows its closeness score within its community. Finally, a grey cell means that the node is not present in the network in that timeframe. The first column from right is TRUE if a node is always leader when it is present in the network and the second column from right counts the number of timeframes where a node is leader.

	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	1953	1519	0	7510	0	5403	13630	0	0	2200	0	0	6
cheryl.johnson@enron.com			4327	0	0	0	7361	0	1559	0	0	0	3
rhonda.denton@enron.com	2037	2260	14108	0	0	0							3
l..denton@enron.com							10885	0	2526	0	3174	0	3
alan.comnes@enron.com									3480	0	0	1926	2
tim.belden@enron.com										3557	0	1882	2
janet.butler@enron.com					8602	5899	0	0	0	0	0	0	2
shelley.corman@enron.com						5184	0	0	0	0	1798	0	2
susan.scott@enron.com				6076	0	0	0	0	0				1
kam.keiser@enron.com								11031	0	0	0	0	1
l..nicolay@enron.com									2677	0	0	0	1
kenneth.lay@enron.com								24608	0	0	0	0	1
d..steffes@enron.com											1237	0	1
veronica.espinoza@enron.com									2368	0	0	0	1
janel.guerrero@enron.com												2119	1
deshonda.hamilton@enron.com			12208	0	0								1
outlook.team@enron.com				4082	0	0	0	0					1
scott.neal@enron.com									3955	0	0	0	1
k..allen@enron.com							9099	0	0	0	0	0	1
stephanie.miller@enron.com							5919	0	0	0	0	0	1
bob.ambrocik@enron.com										1273			1
kimberly.bates@enron.com								4752	0	0	0	0	1
Number of mediators in each timeframe	2	2	3	3	1	3	5	3	6	3	3	3	

Figure 4.12: This figure shows how mediator role in each node changes through time. For each node in each timeframe, there is either a 0, a number greater than 0, or a grey cell. 0 means that the node is present in that timeframe but not as a mediator, the number means the node is mediator and shows its mediator score within the network, and finally a grey cell means that the node is not present in the network in that timeframe.

defined and studied in [67, 69, 68]. In this section, we are interested in finding the possible interaction or mutual relation between the role events and community events in the Enron communication dataset.

Table 4.1 presents important role-community event mappings for community C10T0. This community is present from January 2001 to December 2001. An interesting mapping between role events and community events is shown in the last raw of Table 4.1. Shelley Corman, VP of regulatory affairs who is one of the leaders of C10T0 in timeframe 9, is not leader anymore in timeframe 10. This change in the role of Shelley Corman is concurrent with dissolution of C10T0 in the beginning of timeframe 11. We are not claiming that the change of the role of Shelley Corman leaded to dissolution of C10T0 and it might be a coincidence, but it is an interesting clue for further investigations.

Table 4.2 shows significant role-community event mappings for community C7T0. C7T0 is present over all studied timeframes, i.e. from January 2001 to the end of December 2001. The examples of how a community event affects role

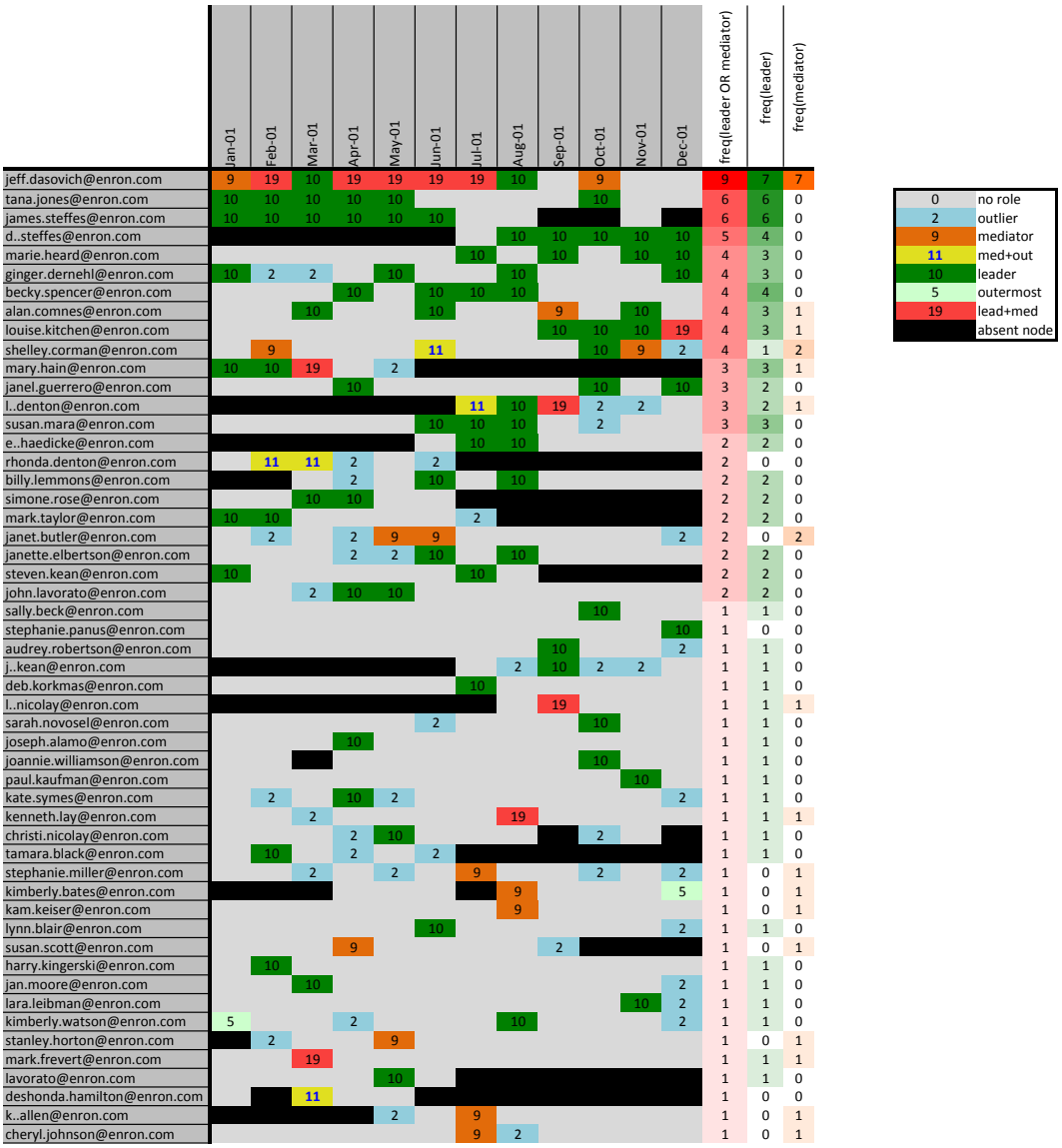


Figure 4.13: In this figure, transitions between different roles in consecutive timeframes are shown. A number is assigned to each role and for nodes having multiple roles, the value of its roles are added. As shown "No role" is represented by 0, outlier by 2, outermost by 5, mediator by 9, and finally leader by 10. These numbers are chosen in a way that the roles of a node is uniquely identified through summation of its roles' assigned numbers. For example, the value 11 for a node infers that it is both mediator and outlier. Different combinations of roles are shown by different colours for the sake of better visualization. Black cells indicate the node is not present in the network in that timeframe and grey cells mean that their associated nodes are present in the network, but do not take a role. In addition, the first, second, and third columns from right respectively count the number of timeframes a node is a mediator, a leader, and the total number of times being either a leader or a mediator.

t	$source(t-1)$	$result(t)$	$event$	$email$	$role_{t-2}$	$role_{t-1}$	$role_t$	com_{t-2}	com_{t-1}	com_t
3	C10T0	C10T0	survive	jan.moore@enron.com	-	leader	-	C10T0	C10T0	C10T0
8	C10T0	C10T0	survive	kimberly.watson@enron.com	-	leader	-	C10T0	C10T0	C10T0
9	C10T0	C10T0	survive	audrey.robertson@enron.com	-	leader	-	C10T0	C10T0	C10T0
11	C10T0		dissolve	shelley.corman@enron.com	leader	mediator	outsider	C10T0	C10T0	-

Table 4.1: Important role-community event mappings for community C10T0 which is present from timeframe 0 (January 2001) to timeframe 11 (December 2001). Null means that the individual is not present in the network at that timeframe, "-" in role columns means that the individual has no role at that timeframe, and in community columns means the individual is not associated with any community at that timeframe.

of a node have happened for Tamara Black who has no role in C7T0, but becomes a leader in C17T1 after C7T0's split into C17T1. There is also Mark Taylor, an employee of Enron, who lost his leading role after the split of C7T0 in timeframe 2. There is a node related to Mark Haedicke, a managing director at Enron, who is not present in the network at timeframe 4, joins community C7T0 at timeframe 5 and at timeframe 6 when a merge happens in C7T0 becomes a leader. Another interesting example is Janette Elbertson, who is an outsider at timeframe 4, joins C7T0 as a leader at timeframe 5, where both a split and a merge happens in C7T0 and stays in the same community without having any role at timeframe 6. Nodes associated with Mark Haedicke and Deb Korkmas become leaders of C7T0 at timeframe 6 which follows by a merge in C7T0 at timeframe 7.

Similar to the above tables, Table 4.3 shows a small portion of role-community event mappings for community C9T0. Interesting changes in roles and communities can be found in this table as well. There are also two interesting dissolve events for communities C17T1 and C20T2. C17T1 is formed at timeframe 1 and dissolves at the beginning of timeframe 3. There is an employee called Tamara Black who is the leader of C17T1 at timeframe 1 and loses her leading role at timeframe 2 followed by C17T1's dissolution at the end of timeframe 2 (beginning of timeframe 3). For C20T2, the life cycle begins at timeframe 2 and the death happens at the beginning of timeframe 9. Rhonda L. Denton, a lawyer, is a leader of C20T2 at timeframe 7 that becomes a mediator together with leader of C20T2 at timeframe 8, and finally becomes an outlier when C20T2 dissolves at the end of timeframe 8.

t	$source(t-1)$	$result(t)$	$event$	$email$	$role_{t-2}$	$role_{t-1}$	$role_t$	com_{t-2}	com_{t-1}	com_t
1	C7T0	C17T1	split	tamara.black@enron.com	null	-	leader	null	C7T0	C17T1
2	C7T0	C7T0	split	mark.taylor@enron.com	leader	leader	-	C7T0	C7T0	C7T0
3	C7T0	C7T0	survive	tana.jones@enron.com	leader	leader	leader	C7T0	C7T0	C7T0
4	C7T0	C7T0	split	becky.spencer@enron.com	-	leader	-	C7T0	C7T0	C7T0
5	C7T0	C7T0	survive	janette.elbertson@enron.com	outsider	outsider	leader	-1	-1	C7T0
6	C7T0	C7T0	merge	e..haedicke@enron.com	null	-	leader	null	C7T0	C7T0
6	C7T0	C7T0	merge	becky.spencer@enron.com	-	leader	leader	C7T0	C7T0	C7T0
6	C7T0	C7T0	merge	janette.elbertson@enron.com	outsider	leader	-	-	C7T0	C7T0
6	C7T0	C11T6	split	marie.heard@enron.com	-	-	leader	C7T0	C7T0	C11T6
6	C7T0	C11T6	split	cheryl.johnson@enron.com	-	-	mediator	C7T0	C7T0	C11T6
6	C7T0	C7T0	split	becky.spencer@enron.com	-	leader	leader	C7T0	C7T0	C7T0
6	C7T0	C7T0	split	janette.elbertson@enron.com	outsider	leader	-	-	C7T0	C7T0
6	C7T0	C7T0	survive	janette.elbertson@enron.com	outsider	leader	-	-	C7T0	C7T0
7	C7T0	C7T0	merge	e..haedicke@enron.com	-	leader	leader	C7T0	C7T0	C7T0
7	C7T0	C7T0	merge	deb.korkmas@enron.com	-	leader	-	C7T0	C7T0	C7T0

Table 4.2: Important role-community event mappings for community C7T0 which is formed at timeframe 0 (January 2001) and never dissolved until the end of timeframe 11 (December 2001). Null means the individual is not present in the network at that timeframe, "-" in role columns means the individual has no role at that timeframe, and in community columns means the individual is not associated with any community at that timeframe.

t	$source(t-1)$	$result(t)$	$event$	$email$	$role_{t-2}$	$role_{t-1}$	$role_t$	com_{t-2}	com_{t-1}	com_t
2	C9T0	C17T2	split	simone.rose@enron.com	-	-	leader	C7T0	C9T0	C17T2
2	C9T0	C17T2	split	mark.frevert@enron.com	-	-	mediator/leader	C9T0	C9T0	C17T2
2	C9T0	C9T0	split/merge	jeff.dasovich@enron.com	mediator	mediator/leader	leader	C9T0	C9T0	C9T0
2	C9T0	C9T0	split/merge	shelley.corman@enron.com	-	mediator	-	C9T0	C9T0	C9T0
2	C9T0	C9T0	split/merge	harry.kingerski@enron.com	-	leader	-	C9T0	C9T0	C9T0
3	C9T0	C9T0	merge	mary.hain@enron.com	leader	mediator/leader	-	C9T0	C9T0	C9T0
3	C9T0	C9T0	merge	alan.comnes@enron.com	-	leader	-	C9T0	C9T0	C9T0
5	C9T0	C9T0	split/merge	ginger.dernehl@enron.com	-	leader	-	C9T0	C9T0	C9T0
5	C9T0	C9T0	split/merge	janet.butler@enron.com	outsider	mediator	mediator	-	C9T0	C9T0
5	C9T0	C9T0	split/merge	christi.nicolay@enron.com	outsider	leader	-	-	C9T0	C9T0
6	C9T0	C9T0	merge	alan.comnes@enron.com	-	leader	-	C9T0	C9T0	C9T0
7	C9T0	C9T0	split/merge	stephanie.miller@enron.com	-	mediator	-	C9T0	C9T0	C9T0
7	C9T0	C9T0	split/merge	steven.kean@enron.com	-	leader	-	C9T0	C9T0	C9T0
7	C9T0	C9T0	split/merge	richard.shapiro@enron.com	-	leader	leader	C9T0	C9T0	C9T0
7	C9T0	C17T2	split	billy.lemmons@enron.com	leader	-	leader	C17T2	C9T0	C17T2
7	C9T0	C17T2	split	k..allen@enron.com	-	mediator	-	C17T2	C9T0	C17T2
8	C9T0	C9T0	merge	jeff.dasovich@enron.com	mediator/leader	leader	-	C9T0	C9T0	C9T0
8	C9T0	C9T0	merge	ginger.dernehl@enron.com	-	leader	-	C9T0	C9T0	C9T0
8	C9T0	C9T0	merge	l..nicolay@enron.com	null	-	mediator/leader	null	C9T0	C9T0
9	C9T0	C7T9	split/merge	jeff.dasovich@enron.com	leader	-	mediator	C9T0	C9T0	C7T9
9	C9T0	C7T9	split/merge	l..nicolay@enron.com	-	mediator/leader	-	C9T0	C9T0	C7T9
9	C9T0	C7T9	split/merge	sarah.novosel@enron.com	-	-	leader	C9T0	C9T0	C7T9
9	C9T0	C7T9	split/merge	alan.comnes@enron.com	-	mediator	-	C9T0	C9T0	C7T9
9	C9T0	C7T9	split/merge	d..steffes@enron.com	leader	leader	leader	C9T0	C9T0	C7T9
9	C9T0	C7T9	split/merge	richard.shapiro@enron.com	leader	leader	leader	C9T0	C9T0	C7T9
9	C9T0	C9T0	split	sally.beck@enron.com	-	-	leader	C17T2	C9T0	C9T0
9	C9T0	C9T0	split	joannie.williamson@enron.com	-	-	leader	C17T2	C9T0	C9T0
9	C9T0	C9T0	split	louise.kitchen@enron.com	-	-	leader	C17T2	C9T0	C9T0

Table 4.3: Important role-community event mappings for community C9T0 which is formed at timeframe 0 (January 2001) and never dissolves until the end of timeframe 11 (December 2001). Null means the individual is not present in the network at that timeframe, "-" in role columns means the individual has no role at that timeframe, and in community columns means the individual is not associated with any community at that timeframe.

4.4 Evaluation and Discussion

In this chapter, we presented our results on applying the structural social role mining framework on the Enron communication dataset. We investigated various metrics to plug into our proposed framework in order to identify the fundamental roles defined as: leader, outermost, and mediator. The results are presented and visualized through descriptive figures. To gain a better understanding of the relation between the identified leaders/mediators (as the prominent roles in the proposed framework) and people involved in the Enron story, Tables 4.4 and 4.5 are presenting their position within the company.

Furthermore, a joint study of the information about the events presented in the timeline of the Enron (Appendix A) and our role mining results could lead to better insights of the events. For instance, in August 2001, Kenneth Lay becomes an important mediator (high mediator score) in the network. Based on the information in the timeline of the Enron events, August is the time when Jeffery Skilling resigns from his position as the CEO and Kenneth Lay becomes the new CEO. Another interesting scenario happens with tracking the changes of Rhonda L. Denton, who is a lawyer according to our findings. She starts being a leader from July, when Jeffery Skilling asked Kenneth Lay to take the position of the CEO after him. Serving as a leader until September, when Rhonda Denton has both important roles, being a leader and a mediator at the same time. Moreover, Skilling sells $15.5 million of stock in September right before Kenneth Lay's report of $618 million loss to the employees in October. The sequence of these events along with the changes of the role of Rhonda Denton especially in September, indicates how they were trying to take an action to reduce their losses before October and especially in September.

There are many investigations on the Enron email network but few of them concentrate on finding important people within this network. Diesner and Carley in [19] compare the changes of the network in 2001 and 2000. Although they build the network in a different way and consider edges to have direction, there are interesting overlaps between our results. Diesner and Carley identify important nodes based on the value of various centrality scores and list the top-5 key-players ac-

cording to each measure. As shown in Figures 4.11 and 4.12 the following are the eleven leaders and the three mediators found by our proposed framework for October 2001: Tana Jones, Richard Shappiro, James Steffes, Marie Heard, Kathryn Sheppard, Louise Kitchen, Sarah Novosel, Joannie Williams, Sally Beck, Holly Keiser, Janel Guerrero, Jeff Dasovich, Tim Belden, and Bib Ambrocik. Comparing to Diesner and Carley's results, there are 3 out of 5 overlaps in the results of each of the categories (closeness, betweenness, eigenvector, and in degree centrality measures) and our identified leaders and mediators. It is worth mentioning that there are 9 people who are identified as leaders/mediators in October 2001 according to our results but not found in the work of Diesner and Carley.

In another work by Shetty and Adibi in [66] influential nodes in the Enron network are identified by considering direct neighbours, neighbours at distance two, highest number of sent emails, and betweenness centrality score. In spite of the differences in building the network graph in our work and theirs, comparisons reveal the following overlaps: Louise Kitchen, Scott Neal, Jeff Dasovich, Kenneth Lay, and Tana Jones as important nodes in both works.

Matching the identified leaders/mediators with their organizational position, comparing these roles with important nodes identified by other works, and using the identified roles along with other information such as temporal role changes and time line events, all together validate and emphasize on the benefits of using our proposed framework in analyzing the underlying network.

email	*position*
tana.jones@enron.com	Senior Legal Specialist
jeff.dasovich@enron.com	Executive/Director for State Government Affairs
james.steffes@enron.com	Vice President
richard.shapiro@enron.com	VP regulatory affairs (Enron'stop lobbyist)
d..steffes@enron.com	*James Steffes*
marie.heard@enron.com	Lawyer
louise.kitchen@enron.com	President of Enron Online
susan.mara@enron.com	California director of Regulatory Affairs
mary.hain@enron.com	In house Lawyer
janel.guerrero@enron.com	An employee in government affairs department
john.lavorato@enron.com	CEO, Enron America
alan.comnes@enron.com	Director Government and Regulatory Affairs
billy.lemmons@enron.com	Vice President
steven.kean@enron.com	VP and Chief of staff
lynn.blair@enron.com	Manager
e..haedicke@enron.com	Managing director, Legal
kate.symes@enron.com	Trader
audrey.robertson@enron.com	Transwestern Pipeline Company

Table 4.4: Leading roles found in Enron communication network and their position in the company.

email	*position*
jeff.dasovich@enron.com	Executive/Director for State Government Affairs
cheryl.johnson@enron.com	Minority Counsel
rhonda.denton@enron.com	Lawyer
l..denton@enron.com	Lawyer
tim.belden@enron.com	Head of Enron's West Coast Trading Desk in Portland Oregon
shelley.corman@enron.com	VP, regulatory affairs
susan.scott@enron.com	Assistant trader
kam.keiser@enron.com	Employee
l..nicolay@enron.com	Senior Director Regulatory Affairs
kenneth.lay@enron.com	CEO, Chairman
d..steffes@enron.com	Vice President
veronica.espinoza@enron.com	Staff of credit risk management
janel.guerrero@enron.com	An employee in government affairs department
outlook.team@enron.com	*mailing list*
scott.neal@enron.com	VP, Trader
k..allen@enron.com	Managing director
bob.ambrocik@enron.co	Manager - Enterprise Storage and Backup Team

Table 4.5: Mediator roles found in Enron communication network and their position in the company.

Chapter 5

CBetweenness and LBetweenness Centrality measures

Betweenness centrality (BC) is a well known centrality measure which was first defined in [27] in 1979 and has been used as a powerful means in many network analysis since then. Modifying Betweenness centrality with the aim of taking into account the structure of communities in a network, led us to the definition of $LBetweenness$ (LBC) and $CBetweenness$ (CBC) centralities in Chapter 3. According to our definitions, only inter-community shortest paths are considered in calculating CBC and LBC.

In this section we present the results of our experiments on comparing BC, LBC, and CBC on the Karate Club network [82] and the Enron communication network described in Chapter 4.

5.1 Karate Club Network

Karate club is the network of friendships between the 34 members of a karate club at a US university. It is a small but informative social network recorded by Zachary from a karate club in a university over the period of three years shown in Figure 5.1. Node 1 and 34 respectively represent the instructor and president of the club. During the three year observation of Zachary a conflict happens between the instructor and the president. As a result, the network splits into supporters of the instructor and supporters of the president.

Table 5.1 presents the top-20 sorted list of nodes in the karate club network

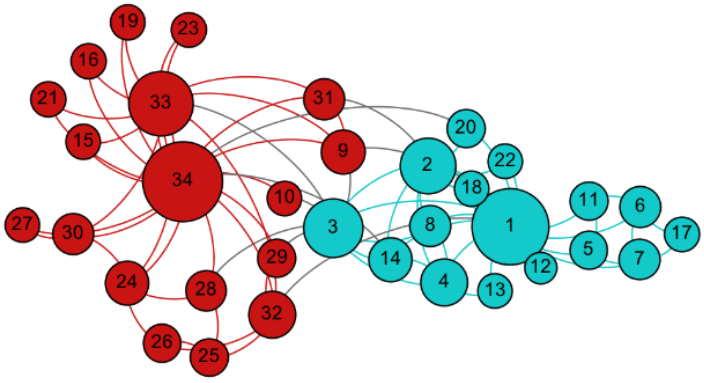


Figure 5.1: Friendship network between 34 karate club members at a university in US.

BC	1	34	33	3	32	9	2	14	20	7	6	28	24	31	4	26	30	25	29	10
CBC	1	3	34	33	14	32	2	28	9	31	7	20	6	4	24	10	30	25	27	22
LBC	3	9	14																	

Table 5.1: Top-20 nodes based on three different criteria: BC, CBC, and LBC scores in the karate club network. Nodes that are shared between BC, CBC, and LBC are colored for better comparison. Interestingly there are only three nodes in the top-20 list for LBC. This happens since there are few leader nodes in this dataset and only the nodes 3,9, and 14 reside on the shortest paths between leaders.

according to their BC, CBC, and LBC scores. The results of the Spearman correlation coefficient between the rankings of all nodes based on these three measures are shown in Table 5.3. There are only three nodes reported for the LBC metric since there are fewer nodes residing on the shortest path between the leaders of communities and in this case there are only three. According to the results, there is a strong correlation between the ordering of nodes based on their BC and CBC in the karate club network. Moreover, these three centrality measures, share a great portion of the nodes in their top-20 lists.

5.2 Enron Communication Network

Table 5.2 presents the top-20 lists for BC, CBC, and LBC for the October 2001 timeframe of the Enron communication network. According to the Table 5.3, there is a high correlation between BC and CBC rankings in the Enron network. How-

BC	CBC	LBC
sally.beck@enron.com	l..denton@enron.com	jeff.dasovich@enron.com
jeff.dasovich@enron.com	jeff.dasovich@enron.com	tim.belden@enron.com
l..denton@enron.com	tim.belden@enron.com	l..denton@enron.com
l..nicolay@enron.com	alan.comnes@enron.com	s..bradford@enron.com
d..steffes@enron.com	l..nicolay@enron.com	mike.grigsby@enron.com
louise.kitchen@enron.com	bob.ambrocik@enron.com	john.lavorato@enron.com
bob.ambrocik@enron.com	louise.kitchen@enron.com	b..sanders@enron.com
richard.shapiro@enron.com	nancy.bagot@enron.com	l..nicolay@enron.com
j..kean@enron.com	d..steffes@enron.com	danny.mccarty@enron.com
tim.belden@enron.com	shelley.corman@enron.com	stanley.horton@enron.com
fraisy.george@enron.com	b..sanders@enron.com	alan.comnes@enron.com
joannie.williamson@enron.com	susan.lindberg@enron.com	fran.chang@enron.com
s..bradford@enron.com	sally.beck@enron.com	drew.fossum@enron.com
nancy.bagot@enron.com	bill.williams@enron.com	brian.redmond@enron.com
b..sanders@enron.com	john.lavorato@enron.com	janette.elbertson@enron.com
veronica.espinoza@enron.com	john.buchanan@enron.com	edward.sacks@enron.com
janel.guerrero@enron.com	janet.butler@enron.com	robert.badeer@enron.com
e..haedicke@enron.com	sarah.novosel@enron.com	jeff.richter@enron.com
shelley.corman@enron.com	j..kean@enron.com	louise.kitchen@enron.com
alan.comnes@enron.com	drew.fossum@enron.com	dave.perrino@enron.com

Table 5.2: Top-20 nodes based on three different criteria: BC, CBC, and LBC scores in the Enron communication network. Nodes that are in more than shared betweene BC and CBC and LBC are colored for better comparison.

ever, the portion of nodes that are being shared among the top-20 lists is less compared to the karate club network. The reason for this phenomena could be caused by the fact that the karate club network consists of only 2 communities, however in the Enron network it is composed of 5 communities. Thus, the effect of communities is more in the Enron communication network and has affected the results.

In consideration of time, Table 5.4 reports the time spent on computing BC[1], CBC, and LBC on the both karate and Enron (Oct. 2001) networks. According to these results, computing LBC takes less time compared to BC and CBC on both networks. Moreover, CBC performs more efficient on the Enron network which is the result of the existence of communities.

[1]Jung implementation is used for computing betweenness centrality. Therefore, it is more optimized than the research code for CBC and LBC that we have implemented.

	BC, CBC	BC, LBC	CBC, LBC
Karate Club	0.94	0.58	0.62
Enron	0.90	0.63	0.72

Table 5.3: Spearman correlation coefficient between Betweenness Centrality (BC), CBetweenness Centrality (CBC), and LBetweenness Centrality (LBC) for the networks of Karate club and Enron's communication network.

	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 5.4: Running time for computing BC, CBC, and LBC on the karate and Enron networks.

5.3 Discussion

In this section we showed how CBC and LBC will result in different ranking lists of nodes compared to BC since these two metrics consider a new parameter that is the community affiliation. Thus, when communities are important factors in determining the centrality of nodes, CBC and LBC are better candidates than BC. Comparing the results within each of the karate and Enron networks as well as comparing the results of these two networks with each other better highlights how the structure of communities can be effective in identifying influential (important) nodes within a network.

From the time complexity point of view, computing CBC and LBC are less expensive than BC. The fact that only inter-community paths are considered is the key to lessen the computational time. However, the extent to which the time complexity is reduced highly depends on the structure of communities and the ratio of paths that reside within communities.

Chapter 6

Conclusion and Discussion

In this thesis we proposed The SSRM framework to study different patterns that nodes may build through their interactions in a social network. To this end, we first defined the concept of *role* in network science based on definitions from social science. Using definition of the role, we introduced our framework. The proposed framework is built based on the assumption that social networks are composed of a set of communities (set of highly connected nodes). Considering the existence of communities, we defined 4 fundamental roles (leader, outermost, mediator, and outsider) in a network. Moreover, we defined new metrics (L-Betweenness (LBC), C-Betweenness (CBC), and DiversityScore (DS)) and proposed general strategies for identifying the aforementioned roles. Finally, we applied the framework to the Enron communication network and observed how the defined roles relate to events happening the network.

The structural social role mining framework presented in this dissertation is inspired by two intrinsic characteristics of humans social life. The first characteristic is the concept of groups as the fundamental components of each society. According to sociology, all societies are composed of multiple groups of people. To put it in other words, people in a society are associated with groups.

The next characteristic of human beings social life, is the role taking behaviour of people in their interactions with friends, family, colleagues, etc. Thus, we defined the concept of social role as the role people take in their society using various definitions of role from social science. Within network science terminology, we defined social role of an entity as its structural and non-structural properties in the

network considering interactions with the rest of the network.

Based on these two prominent assumptions, we proposed the structural social role mining framework. In this framework, we defined 4 fundamental roles named leader, outermost, mediator, and outsider. These roles are either inter-community or intra-community roles. Leader and outermost are intra-community roles as they are defined within a community, while mediator and outsider are defined within the whole network as inter-community roles. More precisely, we defined leaders as the most central/important nodes and outermosts as the least central nodes within each community. Furthermore, we defined outsiders as the nodes that are not affiliated with any community in the network. Finally, mediators are defined as the nodes that play an important role in connecting distinct communities to each other.

In order to identify the set of aforementioned roles, we need a methodology to extract them from the complicated structure of social networks. Identifying outsiders are quite straight forward based on the assumption that social networks are composed of communities. Thus, having communities in a network, nodes that are not a member of any of those communities are identified as outliers. To identify leaders and outermosts within a community, our proposed methodology suggests to choose an appropriate metric (such as centrality measures) and compute a score for each node based on that measure. Then, rank nodes by their computed scores and extract leaders and outermosts by analyzing the sorted list of nodes to find the most and the least significant ones. For identifying mediators, we defined new measures called L-Betweenness (LBC), C-Betweenness (CBC) and DiversityScore (DS). The idea behind the definition of LBC and CBC, is to overcome the time complexity of Betweenness centrality in large social networks. Intuitively, it is not necessary to compute shortest paths between all pairs of nodes in the graph to find how much a node is intervening between others. Therefore, the idea is to consider shortest paths only among a subset of nodes (or important nodes) rather than the whole network. This strategy yields to a small subset of nodes. Hence, LBC and CBC respectively are computed by considering shortest paths between pairs of community leaders and pairs of nodes from distinct communities. Therefore, these two metrics are computed more efficiently in very large graphs. Using these newly

defined measures, we calculate a score for all nodes of the network. Then, sort nodes based on their score and identify mediators by either analyzing the sorted list of nodes or using MedExtractor (Algorithm 1).

The aforementioned roles are defined focusing only on the structural properties of the nodes. Thus, they are called structural social roles. However, non-structural characteristic of nodes can be added to the structural ones in order to enrich the role definition. It should be noted that non-structural characteristic are domain dependent and can only be determined knowing information about the dataset.

Moreover, the roles introduced in our framework are basically defined in a single snapshot of the network when the ranking measures are considering only one timeframe. On the other hand, if temporal centrality measures or any other temporal metric is used, the framework can identify dynamic (temporal) roles. Either way, tracking how these roles change through time provides us with information about the temporal characteristics of nodes and the network. Hence, dynamic roles can also be defined by the temporal patterns of changes in the roles defined in our proposed framework.

To evaluate our proposed 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. Thus, we tried to find information about the people associated with nodes having the role of a leader or a mediator. According to what we found, they are important people in the Enron organizational hierarchy. Moreover, we observed how nodes change their role through time. Based on role changes we analyzed community changes and other events happening in the Enron's timeline.

Using the structural social role mining framework, we can study the concepts of influence, trust, idea innovators, and other roles that are being studied sporadically, all using a unified framework. The important fact about this framework is how to build the graph (network), connections, and nodes in a way that best represents the objective of the problem. The other important step in using this framework is the appropriate choice of metrics for calculating nodes' scores and sorting them accordingly to identify leaders, outermosts, and mediators.

Within the proposed framework, we studied how change of roles could describe other events in the network. However, this framework lacks methodologies to study dynamic roles. To this end, introducing well-defined dynamic roles based on the four fundamental roles (leader, outermost, mediator, outsider) and proposing methods to identify them is a possible work to extend this framework. Furthermore, developing a way to define and identify roles that are investigated in other works using the structural social role mining is a significant step in order to generalize the use of this framework. This way, the structural social role mining framework can be used to unify the study of roles in social networks.

Bibliography

[1] AGARWAL, N., AND LIU, H. Blogosphere: Research issues, tools, and applications. *ACM SIGKDD Explorations Newsletter 10*, 1 (2008), pp. 18–31.

[2] AGARWAL, N., LIU, H., TANG, L., AND YU, P. S. Identifying the influential bloggers in a community. In *Proceedings of the International Conference on Web Search and Data Mining* (2008), ACM, pp. 207–218.

[3] AKHLAGHPOUR, H., GHODSI, M., HAGHPANAH, N., MIRROKNI, V. S., MAHINI, H., AND NIKZAD, A. Optimal iterative pricing over social networks. In *Internet and Network Economics*. Springer, (2010), pp. 415–423.

[4] AKRITIDIS, L., KATSAROS, D., AND BOZANIS, P. Identifying influential bloggers: Time does matter. In *Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies* (2009), vol. 1, pp. 76–83.

[5] ASUR, S., AND HUBERMAN, B. A. Predicting the future with social media. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (2010), vol. 1, pp. 492–499.

[6] BALOG, K., AND DE RIJKE, M. Determining expert profiles (with an application to expert finding). In *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (Heyderabad, India, 2007), vol. 7, pp. 2657–2662.

[7] BARABÁSI, A.-L., AND ALBERT, R. Emergence of scaling in random networks. *science 286*, 5439 (1999), pp. 509–512.

[8] BEKKERMAN, R. Automatic categorization of email into folders: Benchmark experiments on enron and sri corpora. *Technical Report* (2004).

[9] BETTENCOURT, B., AND SHELDON, K. Social roles as mechanism for psychological need satisfaction within social groups. *Journal of personality and social psychology 81*, 6 (2001), pp. 1131.

[10] BIDDLE, B. J. Recent development in role theory. *Annual review of sociology* (1986), pp. 67–92.

[11] BONACICH, P. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology 2*, 1 (1972), pp. 113–120.

[12] BONACICH, P., AND LLOYD, P. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks 23*, 3 (2001), pp. 191 – 201.

[13] BRITANNICA, E. Role.
http://www.britannica.com/EBchecked/topic/507038/role, [Last accessed 19 March 2014].

[14] CASTEIGTS, A., FLOCCHINI, P., QUATTROCIOCCHI, W., AND SANTORO, N. Time-varying graphs and dynamic networks. In *Ad-hoc, Mobile, and Wireless Networks*. Springer, (2011), pp. 346–359.

[15] CHA, M., HADDADI, H., BENEVENUTO, F., AND GUMMADI, P. K. Measuring user influence in twitter: The million follower fallacy. In *Proceedings on the International Conference on Weblogs and Social Media* (2010), vol. 10, pp. 10–17.

[16] CHAKRABARTI, S., DOM, B. E., KUMAR, S. R., RAGHAVAN, P., RAJAGOPALAN, S., TOMKINS, A., GIBSON, D., AND KLEINBERG, J. Mining the web's link structure. *Computer 32*, 8 (1999), pp. 60–67.

[17] CHEN, J., ZAÏANE, O., AND GOEBEL, R. Local community identification in social networks. In *Proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM)* (2009), pp. 237–242.

[18] DE SOLA POOL, I., AND KOCHEN, M. Contacts and influence. *Social networks 1*, 1 (1979), pp. 5–51.

[19] DIESNER, J., AND CARLEY, K. M. Exploration of communication networks from the enron email corpus. In *Proceedings of the SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security* (2005).

[20] DIESNER, J., FRANTZ, T. L., AND CARLEY, K. M. Communication networks from the enron email corpus "it's always about the people. enron is no different". *Computational & Mathematical Organization Theory 11*, 3 (2005), pp. 201–228.

[21] DOM, B., EIRON, I., COZZI, A., AND ZHANG, Y. Graph-based ranking algorithms for e-mail expertise analysis. In *Proceedings of the 8th SIGMOD workshop on Research issues in data mining and knowledge discovery* (2003), ACM, pp. 42–48.

[22] DOMINGOS, P. Mining social networks for viral marketing. *IEEE Intelligent Systems 20*, 1 (2005), pp. 80–82.

[23] DOREIAN, P., BATAGELJ, V., AND FERLIGOJ, A. *Generalized blockmodeling*, vol. 25. Cambridge University Press, 2005.

[24] FORESTIER, M., STAVRIANOU, A., VELCIN, J., AND ZIGHED, D. A. Roles in social networks: Methodologies and research issues. *Web Intelligence and Agent Systems 10*, 1 (2012), pp. 117–133.

[25] FORTUNATO, S. Community detection in graphs. *Physics Reports 486*, 3 (2010), pp. 75–174.

[26] FOX, L. *Enron: The rise and fall*. John Wiley and Sons, 2003.

[27] FREEMAN, L. C. Centrality in social networks conceptual clarification. *Social networks 1*, 3 (1979), pp. 215–239.

[28] FRIEDMAN, W. *About time: Inventing the fourth dimension*. The MIT press, 1990.

[29] GOFFMAN, E. The presentation of self in everyday life. *New York* (1959).

[30] GOLDER, S. A., AND DONATH, J. Social roles in electronic communities. *Internet Research 5* (2004), pp. 19–22.

[31] HARTLINE, J., MIRROKNI, V., AND SUNDARARAJAN, M. Optimal marketing strategies over social networks. In *Proceedings of the 17th international conference on World Wide Web* (2008), pp. 189–198.

[32] KAPLAN, A. M., AND HAENLEIN, M. Users of the world, unite! the challenges and opportunities of social media. *Business horizons 53*, 1 (2010), pp. 59–68.

[33] KATZ, L. A new status index derived from sociometric analysis. *Psychometrika 18*, 1 (1953), pp. 39–43.

[34] KELLER, E., AND BERRY, J. *The Influentials: One American in ten tells the other nine how to vote, where to eat, and what to buy*. Simon and Schuster, 2003.

[35] KELLY, J. A., ST LAWRENCE, J. S., DIAZ, Y. E., STEVENSON, L. Y., HAUTH, A. C., BRASFIELD, T. L., KALICHMAN, S. C., SMITH, J. E., AND ANDREW, M. E. Hiv risk behavior reduction following intervention with key opinion leaders of population: an experimental analysis. *American Journal of Public Health 81*, 2 (1991), pp. 168–171.

[36] KEMPE, D., KLEINBERG, J., AND TARDOS, É. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (2003), pp. 137–146.

[37] KEMPE, D., KLEINBERG, J., AND TARDOS, É. Influential nodes in a diffusion model for social networks. In *Automata, languages and programming*. Springer, (2005), pp. 1127–1138.

[38] KIETZMANN, J. H., HERMKENS, K., MCCARTHY, I. P., AND SILVESTRE, B. S. Social media? get serious! understanding the functional building blocks of social media. *Business Horizons 54*, 3 (2011), pp. 241–251.

[39] KIM, E. S., AND HAN, S. S. An analytical way to find influencers on social networks and validate their effects in disseminating social games. In *Proceedings of the International Conference on Advances in Social Network Analysis and Mining* (2009), pp. 41–46.

[40] KLIMT, B., AND YANG, Y. The enron corpus: A new dataset for email classification research. In *Proceedings of the European Conference on Machine Learning*. Springer, (2004), pp. 217–226.

[41] KOSSINETS, G., KLEINBERG, J., AND WATTS, D. The structure of information pathways in a social communication network. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (2008), pp. 435–443.

[42] KOSTAKOS, V. Temporal graphs. *Physica A: Statistical Mechanics and its Applications 388*, 6 (2009), pp. 1007–1023.

[43] KWAK, H., LEE, C., PARK, H., AND MOON, S. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World Wide Web* (2010), pp. 591–600.

[44] LESKOVEC, J., ADAMIC, L. A., AND HUBERMAN, B. A. The dynamics of viral marketing. *ACM Transactions on the Web 1*, 1 (2007), pp. 5.

[45] LOMAS, J., ENKIN, M., ANDERSON, G. M., HANNAH, W. J., VAYDA, E., AND SINGER, J. Opinion leaders vs audit and feedback to implement practice guidelines. *The journal of the American Medical Association 265*, 17 (1991), pp. 2202–2207.

[46] MARKS, R. Enron Timeline 2001. http://www.agsm.edu.au/bobm/teaching/BE/Enron/timeline.html, [Last accessed 21 March 2014].

[47] MCCALLUM, A., WANG, X., AND CORRADA-EMMANUEL, A. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artifitial Intelligence Research (JAIR) 30* (2007), pp. 249–272.

[48] MCLEAN, B., AND ELKIND, P. The smartest guys in the room: The amazing rise and scandalous fall of enron. *New York* (2004).

[49] MILGRAM, S. The small world problem. *Psychology today 2*, 1 (1967), pp. 60–67.

[50] NEWMAN, M. E. Fast algorithm for detecting community structure in networks. *Physical review E 69*, 6 (2004), pp. 066133.

[51] OESER, O., AND HARARY, F. A mathematical model for structural role theory: Ii. *Human Relations* (1964).

[52] OESER, O., AND O'BRIEN, G. A mathematical model for structural role theory: Iii. *Human Relations* (1967).

[53] OESER, O. A., AND HARARY, F. A mathematical model for structural role theory: I. *Human Relations* (1962).

[54] PAGE, L., BRIN, S., MOTWANI, R., AND WINOGRAD, T. The pagerank citation ranking: bringing order to the web. *Technical report* (1998).

[55] PALLA, G., DERÉNYI, I., FARKAS, I., AND VICSEK, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature 435*, 7043 (2005), pp. 814–818.

[56] PONS, P., AND LATAPY, M. Computing communities in large networks using random walks. *Computer and Information Sciences* (2005), pp. 284–293.

[57] RABBANY, R., CHEN, J., AND ZAÏANE, O. R. Top leaders community detection approach in information networks. In *Proceedings of the 4th Workshop on Social Network Mining and Analysis* (2010).

[58] RUHNAU, B. Eigenvector-centrality – a node-centrality? *Social networks 22*, 4 (2000), pp. 357–365.

[59] SAKAKI, T., OKAZAKI, M., AND MATSUO, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web* (2010), pp. 851–860.

[60] SANTORO, N., QUATTROCIOCCHI, W., FLOCCHINI, P., CASTEIGTS, A., AND AMBLARD, F. Time-varying graphs and social network analysis: Temporal indicators and metrics. *arXiv preprint arXiv:1102.0629* (2011).

[61] SCOTT, J., AND CARRINGTON, P. J. *The SAGE handbook of social network analysis*. SAGE publications, 2011.

[62] SCRIPPS, J., TAN, P.-N., AND ESFAHANIAN, A.-H. Node roles and community structure in networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (San Jose, California, 2007), pp. 26–35.

[63] SEARCH, O. T. ENRON Minutes for August. http://datasets.opentestset.com/datasets/Enron_files/full/watson-k, [Last accessed 19 March 2014].

[64] SEARCH, O. T. ENRON Minutes for July. http://datasets.opentestset.com/datasets/Enron_files/full/watson-k, [Last accessed 19 March 2014].

[65] SHETTY, J., AND ADIBI, J. The enron email dataset database schema and brief statistical report. *Information sciences institute technical report 4* (2004).

[66] SHETTY, J., AND ADIBI, J. Discovering important nodes through graph entropy the case of enron email database. In *Proceedings of the 3rd international workshop on Link discovery* (2005), pp. 74–81.

[67] TAKAFFOLI, M., FAGNAN, J., SANGI, F., AND ZAÏANE, O. R. Tracking changes in dynamic information networks. In *Proceedings of the International Conference on Computational Aspects of Social Networks* (2011), pp. 94–101.

[68] TAKAFFOLI, M., RABBANY, R., AND ZAÏANE, R. R. Incremental local community identification in dynamic social networks. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining* (August 2013).

[69] TAKAFFOLI, M., SANGI, F., FAGNAN, J., AND ZAÏANE, O. R. Community evolution mining in dynamic social networks. *Procedia-Social and Behavioral Sciences 22* (2011), pp. 49–58.

[70] TAKAFFOLI, M., SANGI, F., FAGNAN, J., AND ZAÏANE, O. R. Modec-modeling and detecting evolutions of communities. In *Proceedings of the International Conference on Weblogs and Social Media* (2011).

[71] TANG, J., MUSOLESI, M., MASCOLO, C., LATORA, V., AND NICOSIA, V. Analysing information flows and key mediators through temporal centrality metrics. In *Proceedings of the 3rd Workshop on Social Network Systems* (2010), p. 3.

[72] TANTIPATHANANANDH, C., BERGER-WOLF, T., AND KEMPE, D. A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (2007), pp. 717–726.

[73] TRAVERS, J., AND MILGRAM, S. An experimental study of the small world problem. *Sociometry* (1969), pp. 425–443.

[74] VALENTE, T. W., AND DAVIS, R. L. Accelerating the diffusion of innovations using opinion leaders. *The ANNALS of the American Academy of Political and Social Science 566*, 1 (November 1999), pp. 55–67.

[75] WATTS, D. J., AND STROGATZ, S. H. Collective dynamics of 'small-world' networks. *nature 393*, 6684 (1998), pp. 440–442.

[76] WEBER, T. Why companies watch your every facebook, youtube, twitter move.
http://www.bbc.co.uk/news/business-11450923, [Last accessed 19 March 2014].

[77] WIKIPEDIA. Enron. http://en.wikipedia.org/wiki/Enron, [Last accessed 19 March 2014].

[78] WIKIPEDIA. Social networks. http://en.wikipedia.org/wiki/Social_network, [Last accessed 19 March 2014].

[79] WILSON, G., AND BANZHAF, W. Discovery of email communication networks from the enron corpus with a genetic algorithm using social network analysis. In *Proceedings of the IEEE Congress on Evolutionary Computation* (2009), pp. 3256–3263.

[80] XU, X., YURUK, N., FENG, Z., AND SCHWEIGER, T. A. Scan: a structural clustering algorithm for networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (2007), pp. 824–833.

[81] XUAN, B. B., FERREIRA, A., AND JARRY, A. Computing shortest, fastest, and foremost journeys in dynamic networks. *International Journal of Foundations of Computer Science 14*, 02 (2003), pp. 267–285.

[82] ZACHARY, W. An information flow model for conflict and fission in small groups. *Journal of anthropological research 33*, 4 (1977), pp. 452–473.

[83] ZHANG, J., ACKERMAN, M. S., AND ADAMIC, L. Expertise networks in online communities: Structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web* (Banff, Alberta, Canada, 2007), pp. 221–230.

Appendix A

A.1 Enron Timeline in 2001

The timeline of events at Enron Corp. in the year 2001 that is of the interest to this work is quoted from [46] as follows:

- **End of 2000** Enron uses aggressive accounting to declare $53 million in earnings for Broadband on a collapsing deal that hadn't earned a penny in profit.

- **Jan. 2001** Belden's West Coast power desk has its most profitable month ever – $254 million in gross profits.
  Jan. 22, 2001 Quarterly Analyst Conference Call – Jeffery Skilling presents a positive report of the company.
  Jan. 25, 2001 Analyst Conference in Houston, Texas. Skilling bullish on the company. Analysts are all convinced. Ken Rice increases his estimates for value of Broadband.

- **Feb., 2001** Tom White resigns from EES (Enron Energy Services, the retail division he headed since 1998) and becomes Secretary of the Army. He cashes out with $14 million and begins to build a huge home in Naples, Florida. The purchase price for the property is $6.5 million.
  Feb., 2001 Over the year 2000 (while he presided over EBS, Enron Broadband Services), Ken Rice cashes in $53 million in shares and options.
  Feb., 2001 Lay retires as CEO and is replaced by Skilling.
  Feb. 5-14, 2001 Senior Andersen partners meet to discuss whether to retain Enron as a client. They call use of mark-to-market accounting "intelligent

gambling."

Feb. 14, 2001 Writer Bethany McLean interviews Skilling.

Feb. 15, 2001 Mark Palmer, head of publicity for Enron, and Fastow go to Fortune to answer questions. Fastow to Bethany McLean: "I don't care what you say about the company. Just don't make me look bad."

Feb. 19, 2001 Fortune article, by Bethany McLean: "Is Enron Overpriced?"

Feb. 21, 2001 Employee Meeting. Skilling says: "Yes, it is a black box. But it is a black box that's growing the wholesale business by about 50 percent in volume and profitability. That's a good black box." Skilling announces Enron's goal: "The World's Leading Company."

- **March, 2001** Enron transfers large portions of EES business into wholesale to hide EES losses.

  March, 2001 Arthur Andersen takes auditor Carl Bass off the Enron account.

  March 23, 2001 Enron schedules unusual analyst conference call to boost stock. It works.

- **April 17, 2001** Quarterly Conference Call.

- **May 17, 2001** Secret meeting at Peninsula Hotel in LA.

- **June 2001** FERC finally institutes price caps across the western states. The California energy crisis ends.

- **July 12, 2001** Quarterly Conference Call. Skilling still presenting the state of the company as awesome and great.

  July 13, 2001 Jeffery Skilling announces desire to resign to Kenneth Lay. Kenneth Lay asks Skilling to take the weekend and think it over. There are two different views of what happened that day. According to Lay, he tried to talk Skilling out of resigning. Skilling says Lay didn't seem to care and that he offered to stay on for six months. Board member says he recommended the transition period to Lay. Lay claims Skilling wanted an immediate out.

  July 24-25, 2001 Skilling meets with analysts and investors in NY. "We will hit those numbers. We will beat those numbers."

- **August 3, 2001** Skilling makes a bullish speech on EES. That afternoon, he lays off 300 employees.

  August 11, 2001 Skilling talks to Mark Palmer about preparing press release for resignation.

  August 13, 2001 Board Meeting. Rick Buy outlines disaster scenario if Enron's stock starts to fall. All SPEs crash. Skilling dismisses this. That evening, in board only session, Skilling, in tears, resigns.

  August 14, 2001 Skilling's Resignation Announcement. In evening, analyst and investor conference call. Skilling: "The company is in great shape" Lay: "Company is in the strongest shape that it's ever been in." Lay is named CEO.

  August 15, 2001 Jim Chanos thinks the stock is going through the floor and bets aggressively on that. Notes that Skilling's departure coincided with release of second quarter 10-Q. Enron's cash flow was a negative $1.3 billion for the first six months.

  Sherron Watkins, an Enron vice president, writes to Lay expressing concerns about Enron's accounting practices.

  August 22, 2001 Ms Watkins meets with Lay and gives him a letter in which she says that Enron might be an "elaborate hoax."

- **September 2001** Skilling sells $15.5 million of stock, bringing stock sales since May 2000 to over $70 million.

  Sept. 26, 2001 Employee Meeting. Lay tells employees: Enron stock is an "incredible bargain." "Third quarter is looking great."

- **Oct. 16, 2001** Enron reports a $618 million third-quarter loss and declares a $1.01 billion non-recurring charge against its balance sheet, partly related to "structured finance" operations run by chief financial officer Andrew Fastow. In the analyst conference call that day, Lay also announces a $1.2 billion cut in shareholder equity.

  Oct. 17, 2001 Wall Street Journal article, written by John Emshwiller and Rebecca Smith, appears. The article reveals, for the first time, the details of Fastow's partnerships and shows the precarious nature of Enron's business.

The SEC begins an informal probe of Enron.

Oct. 22, 2001 Enron acknowledges Securities and Exchange Commission inquiry into a possible conflict of interest related to the company's dealings with the partnerships.

Oct. 23, 2001 Lay professes support for Fastow, saying he has the "highest regard" for his character during conference call with analysts, and employee meeting: "Andy has operated in the most ethical and appropriate manner possible."

Oct. 23, 2001 In a massive shredding operation, Arthur Andersen destroys one ton of Enron documents.

Oct. 24, 2001 Enron ousts Fastow.

Oct. 26-29, 2001 In vain Lay calls top government officials to solicit help for Enron, including Alan Greenspan, Paul O'Neill, and Donald Evans, respectively the chairman of the Fed, the Treasury secretary, and the commerce secretary.

Oct. 31, 2001 Enron announces the SEC inquiry has been upgraded to a formal investigation.

- **Nov. 8, 2001** Enron files documents with SEC revising its financial statements for past five years to account for $586 million in losses. The company starts negotiations to sell itself to Dynegy, a smaller rival, to head off bankruptcy.

  Nov. 9, 2001 Dynegy agrees to buy Enron for about $9 billion in stock and cash.

  Nov. 19, 2001 Enron restates its third quarter earnings and discloses it is trying to restructure a $690 million obligation that could come due Nov. 27.

  Nov. 28, 2001 Enron shares plunge below $1.

  Nov. 29, 2001 Dynegy withdraws from the deal.

- **Dec. 2, 2001** Enron files for Chapter 11 bankruptcy protection, at the time the largest bankruptcy in US history.

Appendix B

B.1 Degree Centrality Distributions

According to the assumption that societies are composed of communities, we extract communities of the Enron communication network in each timeframe. In this section, we show the degree distributions for communities in each timeframe. To ignore very sparse distributions, we only considered communities with more than 25 nodes. Depending on the size of the communities in each, there are timeframes with 2, 3 or 4 communities of sizes larger than 25.

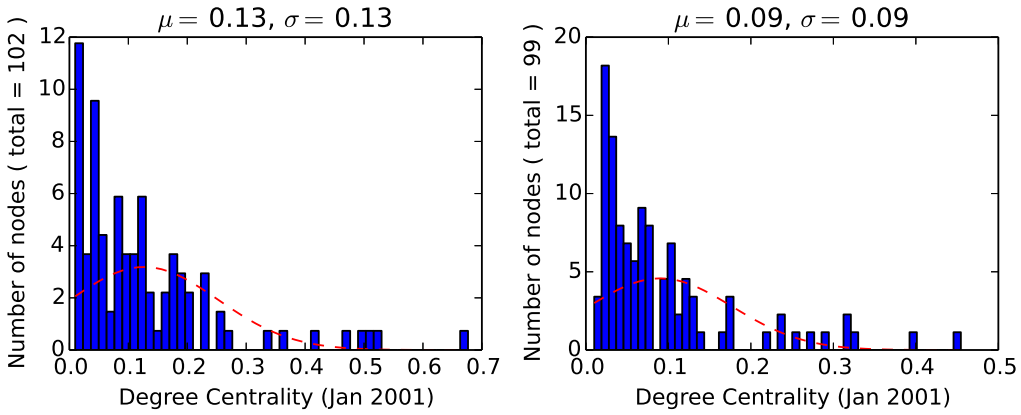


Figure B.1: Plot of Communities' Degree Distribution for January 2001.

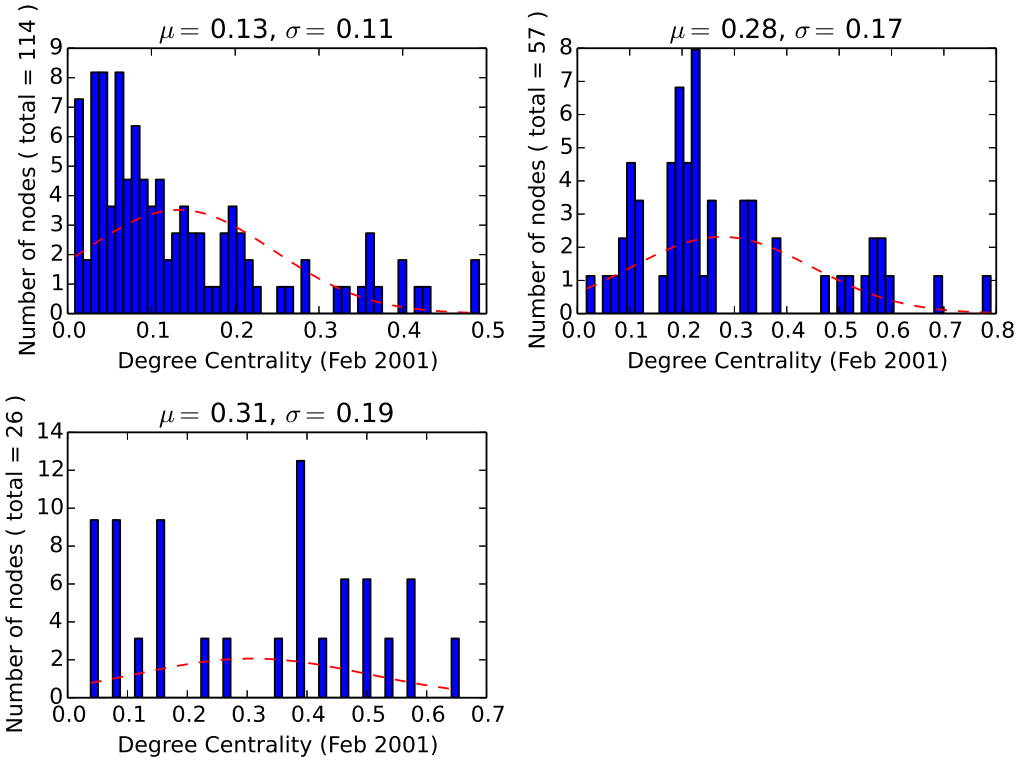


Figure B.2: Plot of Communities' Degree Distribution for February 2001.

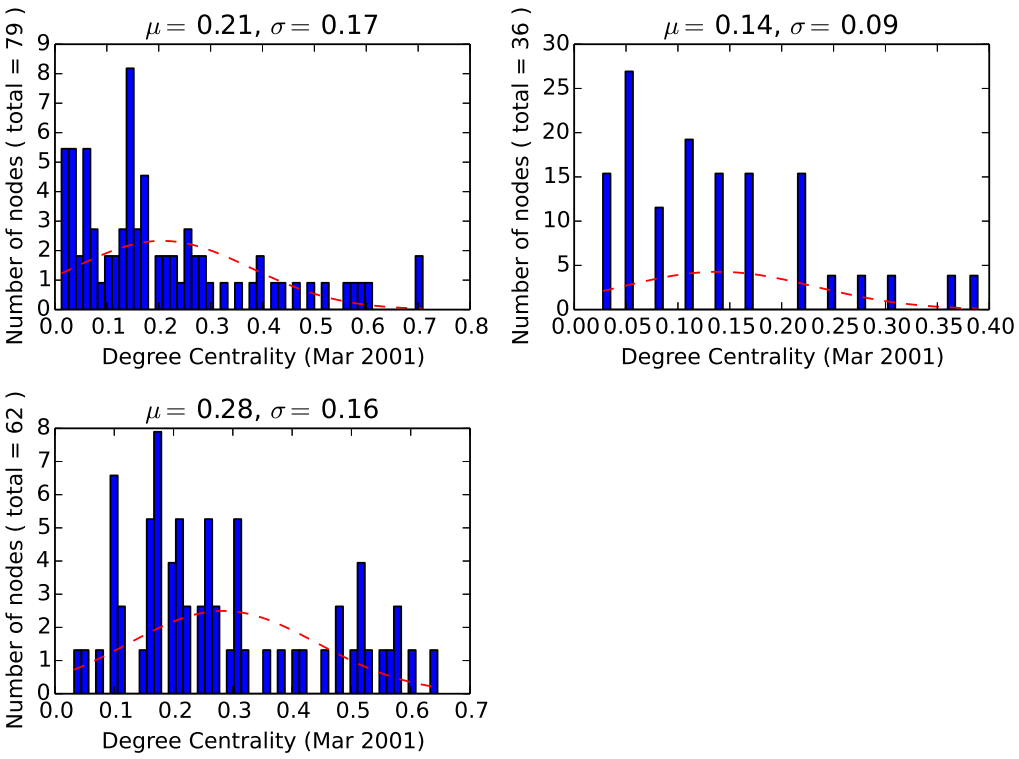


Figure B.3: Plot of Communities' Degree Distribution for March 2001.

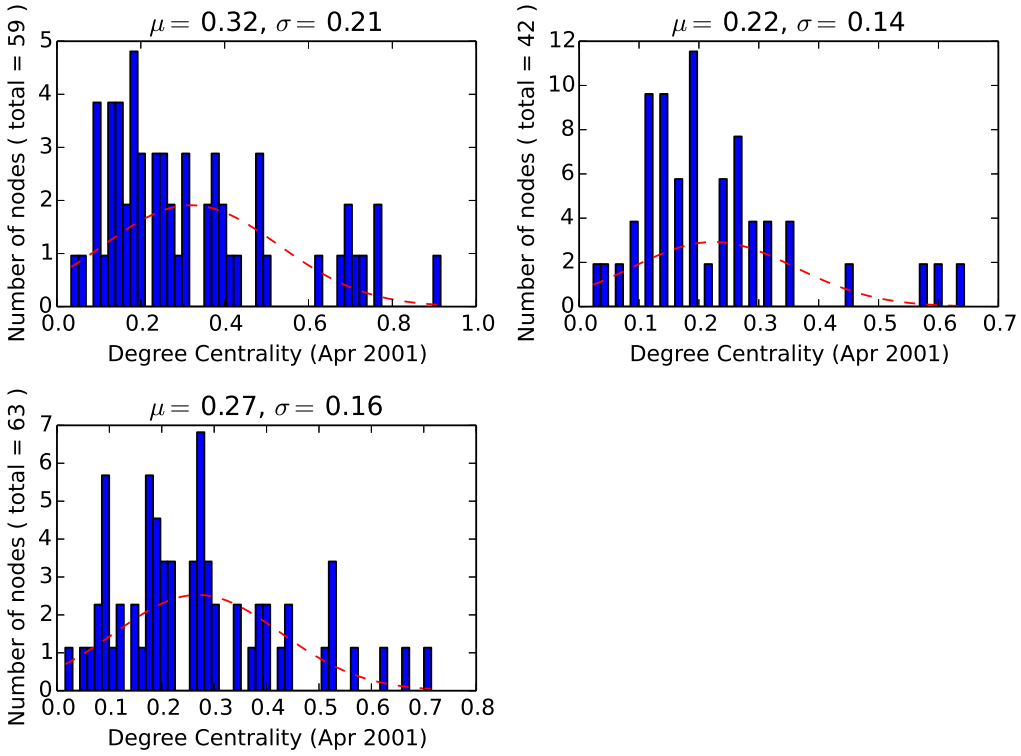


Figure B.4: Plot of Communities' Degree Distribution for April 2001.

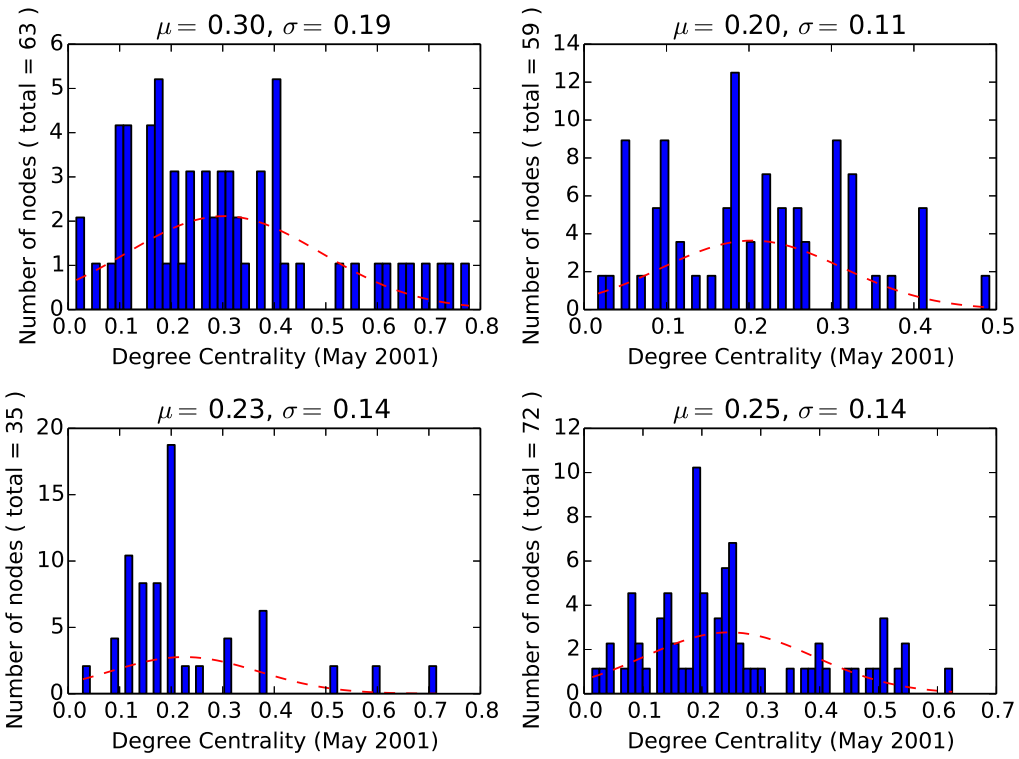


Figure B.5: Plot of Communities' Degree Distribution for May 2001.

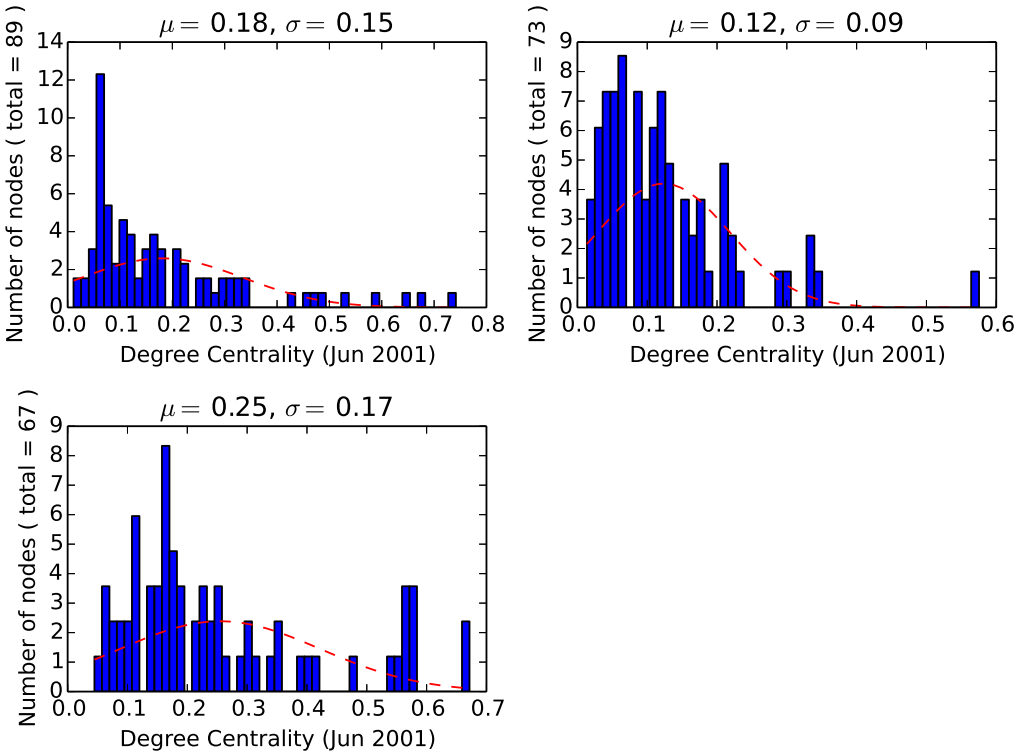


Figure B.6: Plot of Communities' Degree Distribution for June 2001.

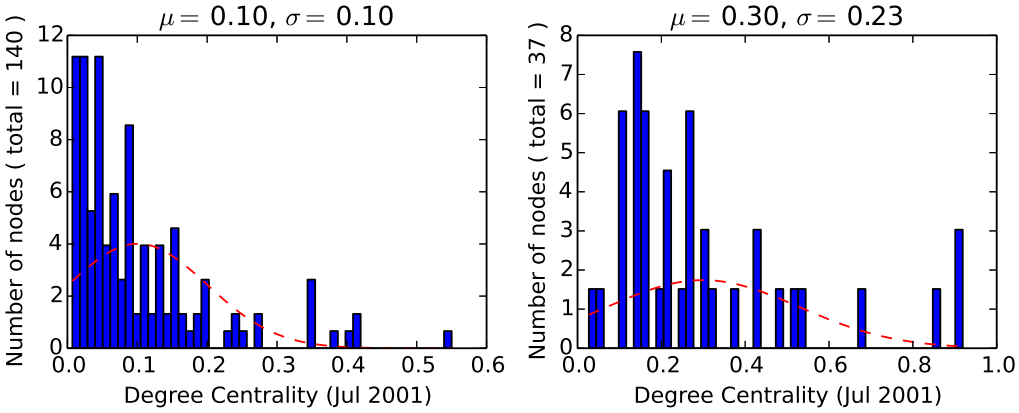


Figure B.7: Plot of Communities' Degree Distribution for July 2001.

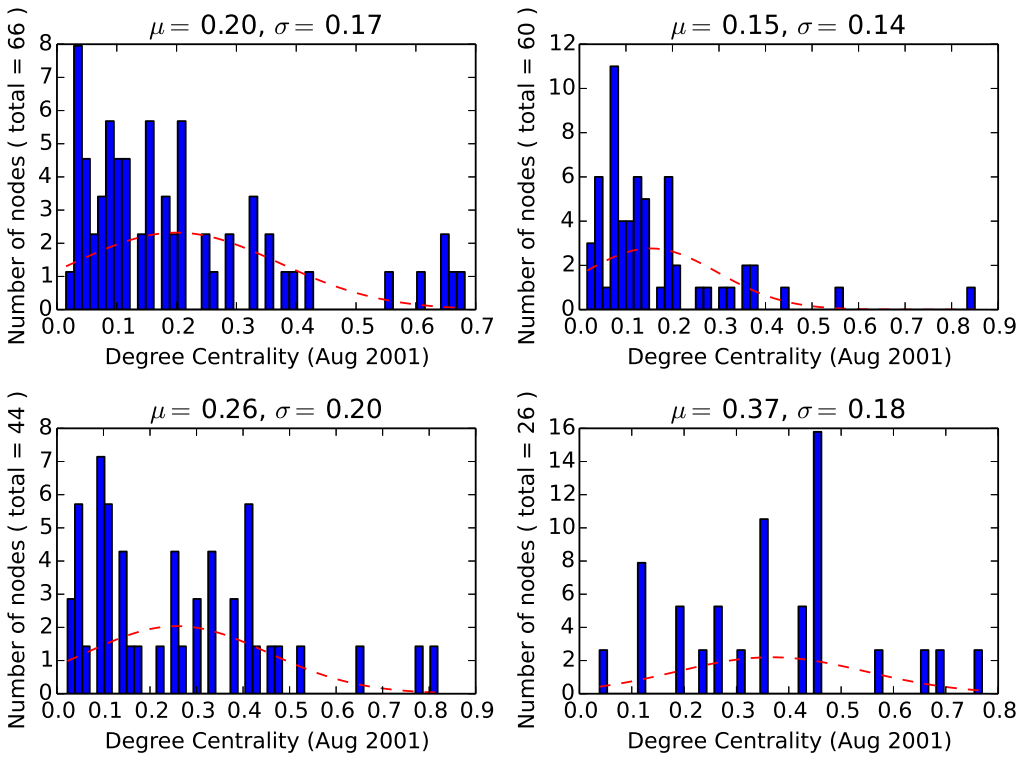


Figure B.8: Plot of Communities' Degree Distribution for August 2001.

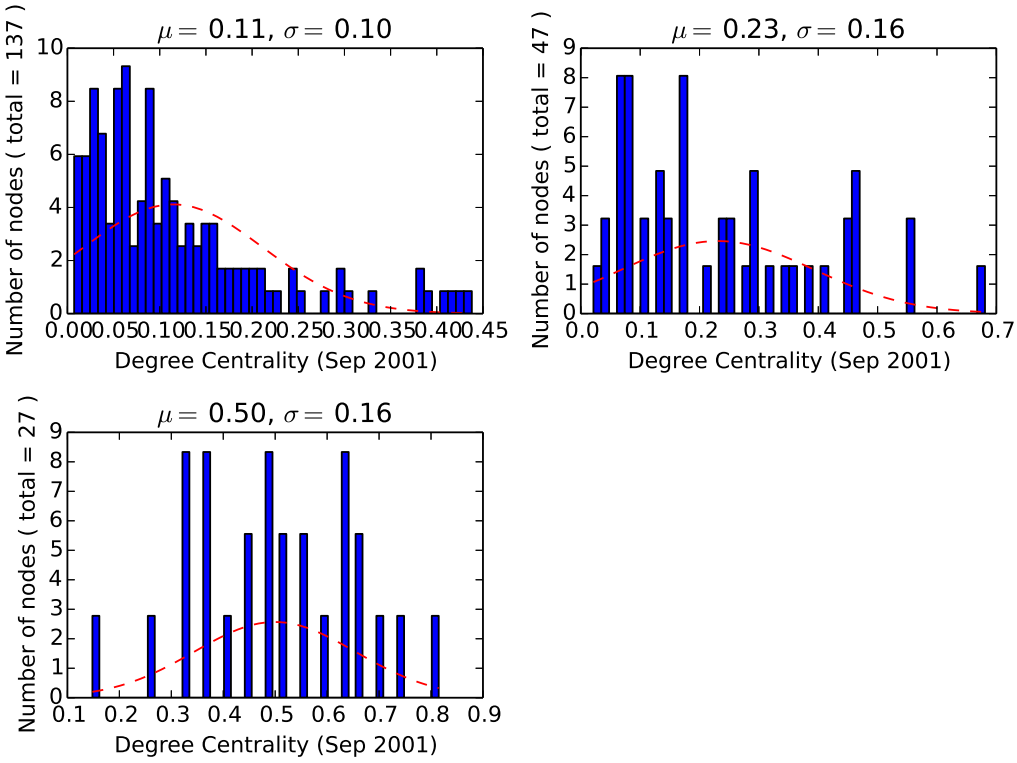


Figure B.9: Plot of Communities' Degree Distribution for September 2001.

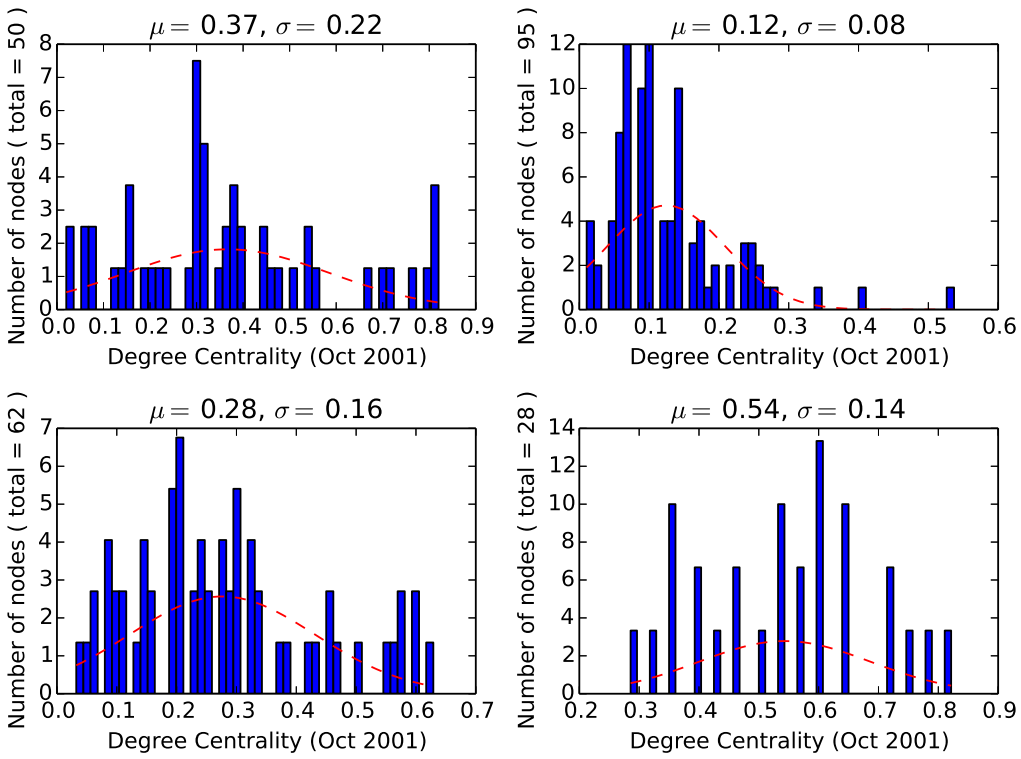


Figure B.10: Plot of Communities' Degree Distribution for October 2001.

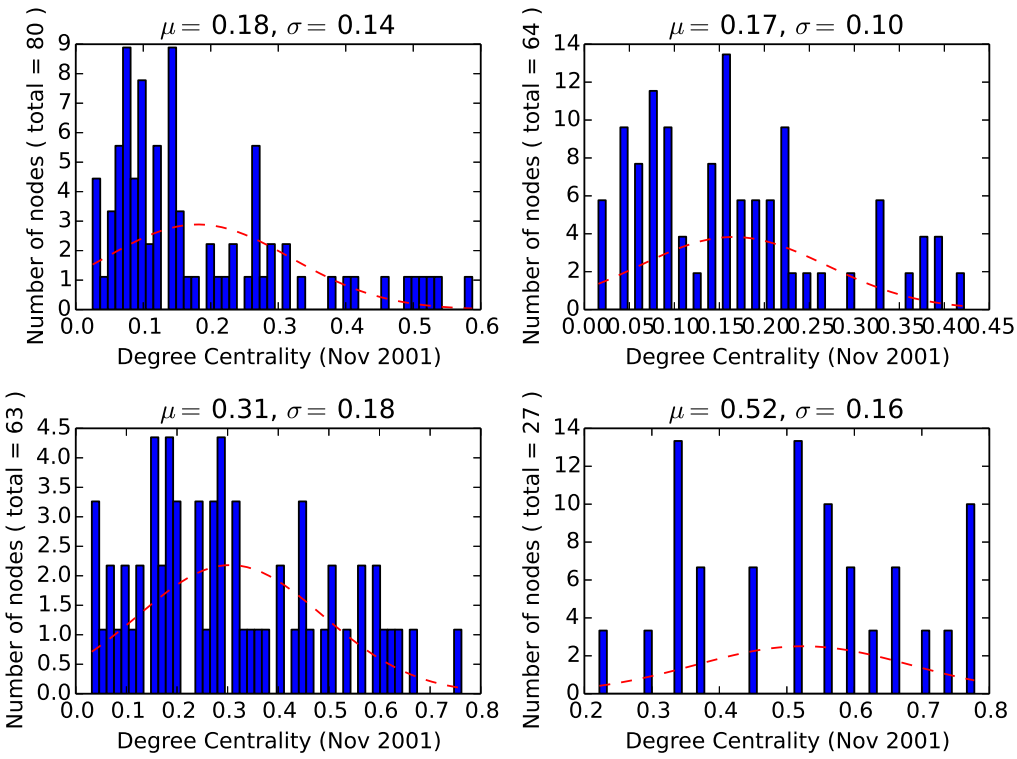


Figure B.11: Plot of Communities' Degree Distribution for November 2001.

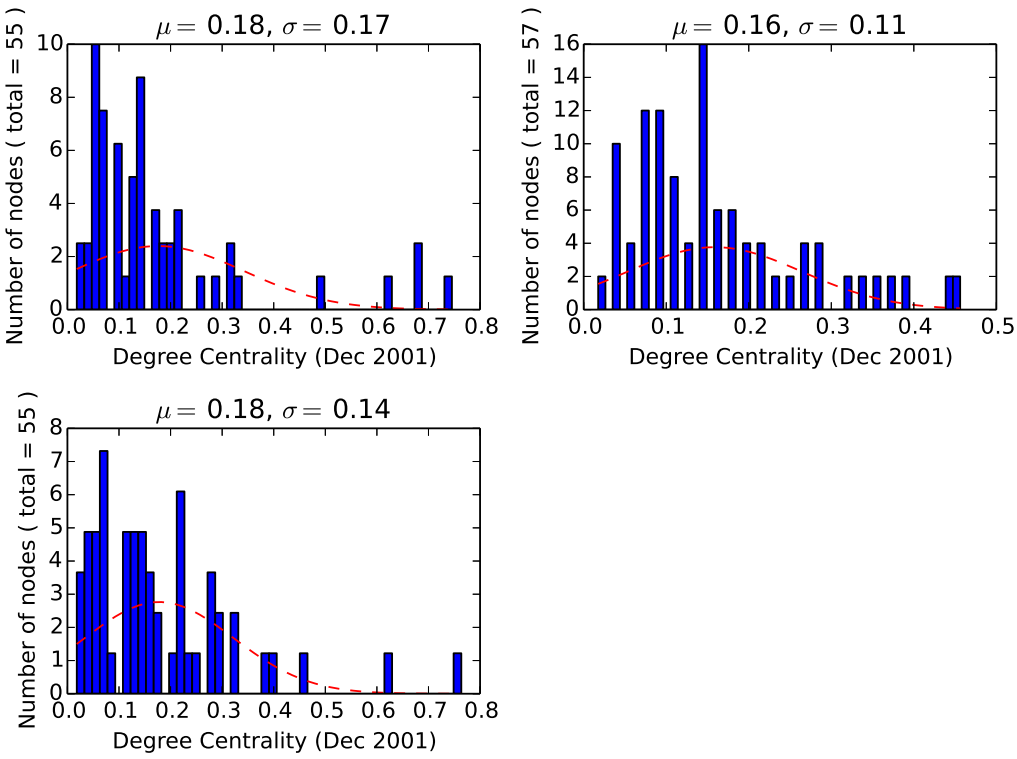


Figure B.12: Plot of Communities' Degree Distribution for December 2001.

B.2 Closeness Centrality Distributions

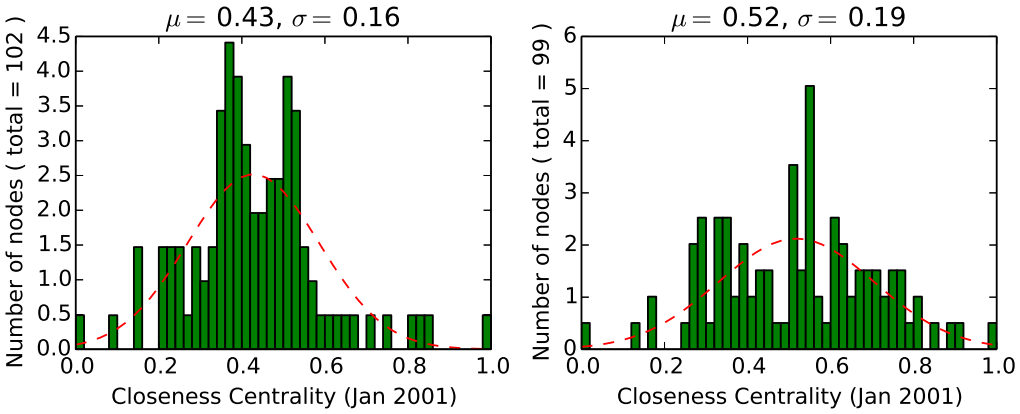


Figure B.13: Plot of Communities' Closeness Distribution for January 2001.

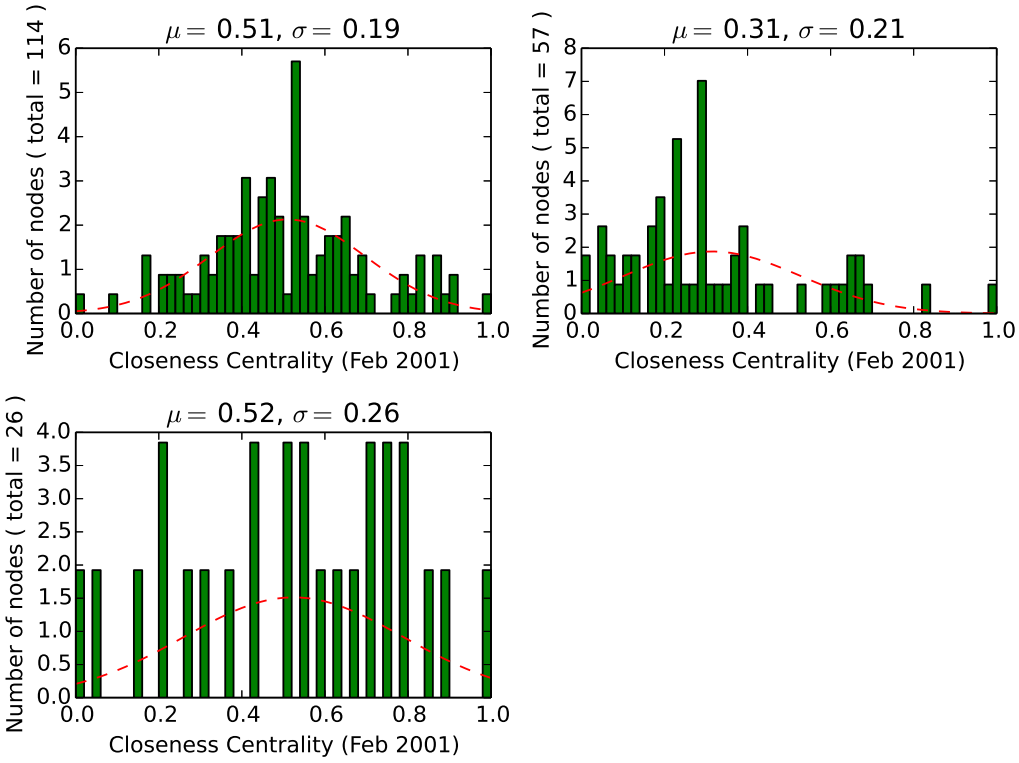


Figure B.14: Plot of Communities' Closeness Distribution for February 2001.

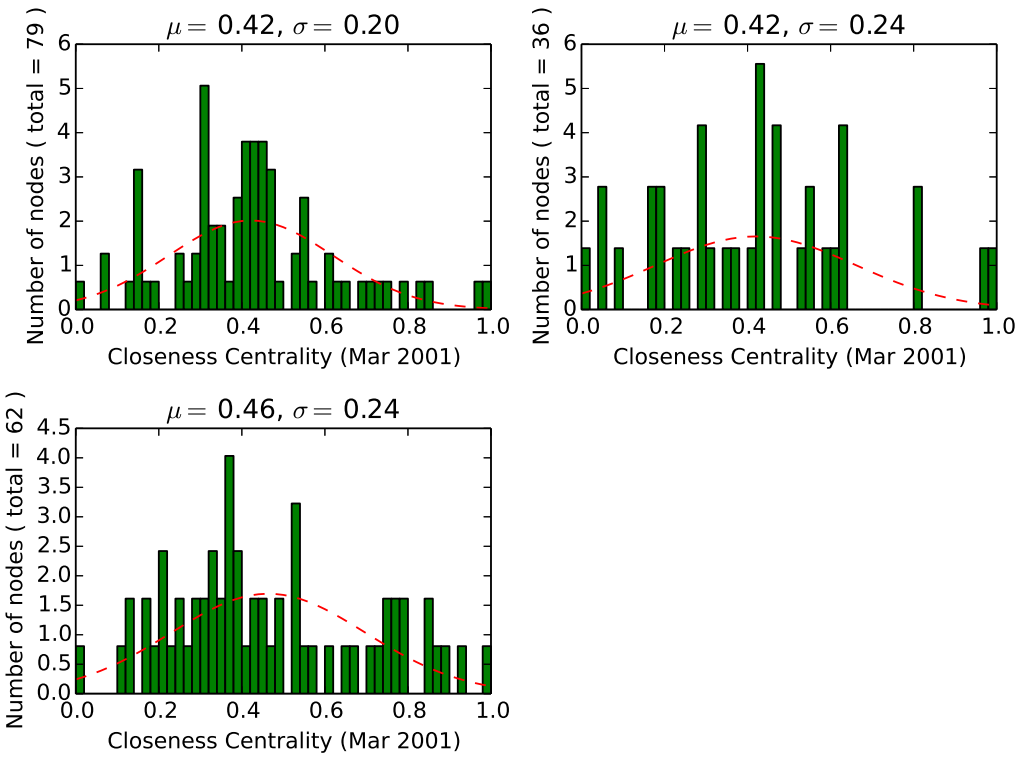


Figure B.15: Plot of Communities' Closeness Distribution for March 2001.

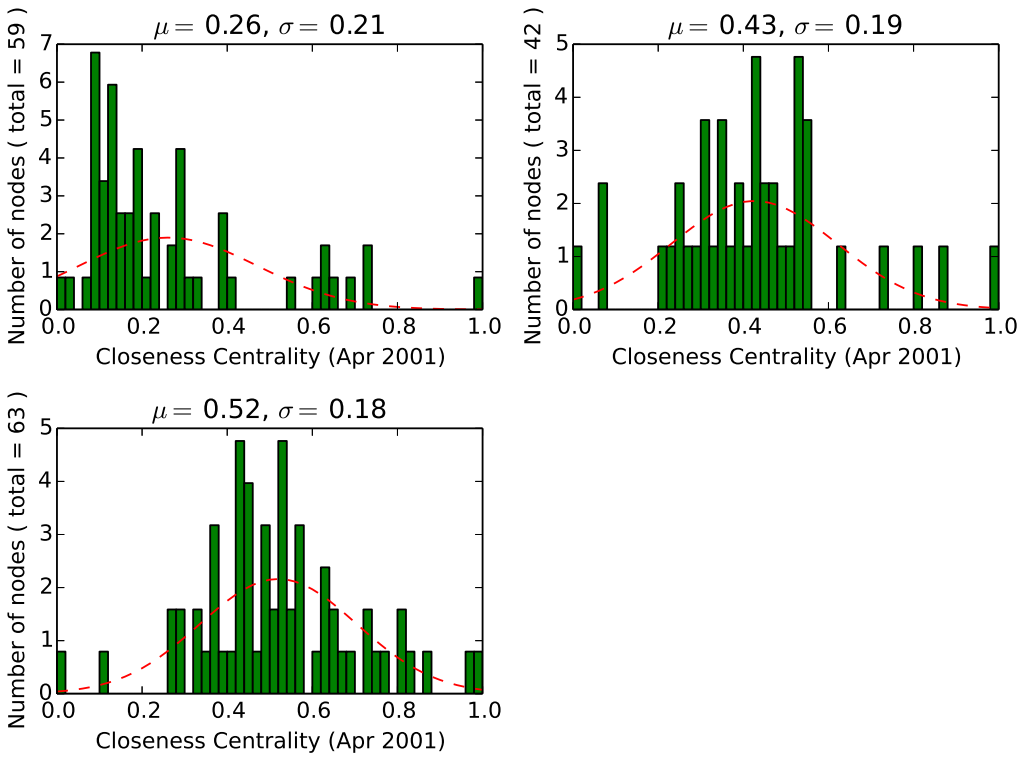


Figure B.16: Plot of Communities' Closeness Distribution for April 2001.

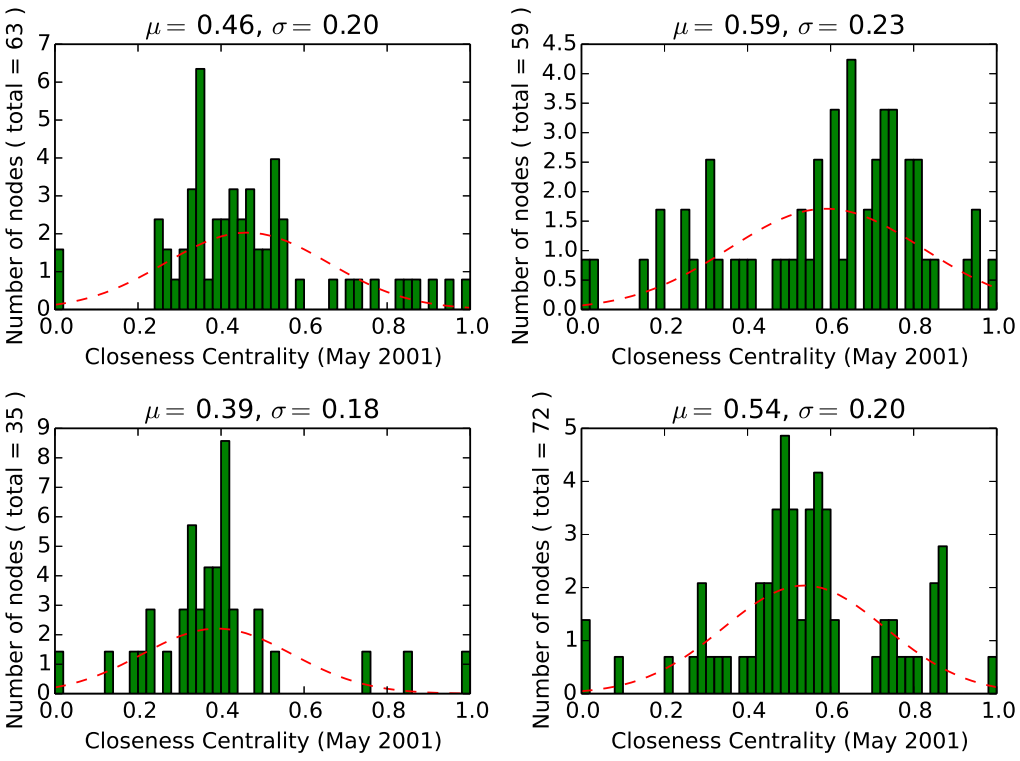


Figure B.17: Plot of Communities' Closeness Distribution for May 2001.

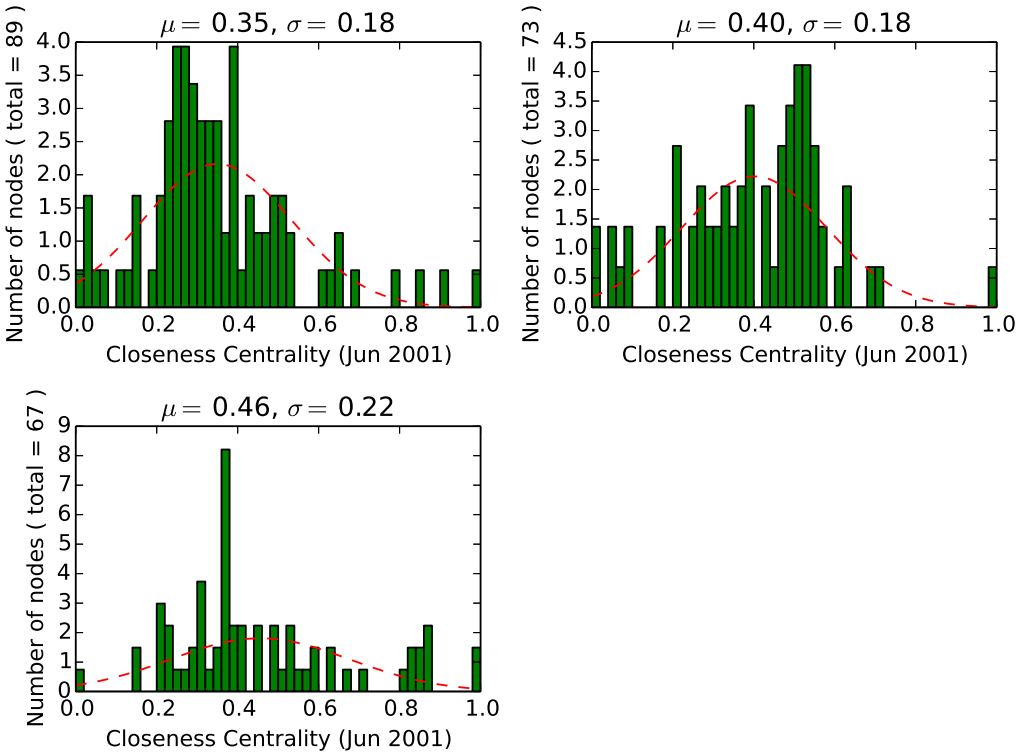


Figure B.18: Plot of Communities' Closeness Distribution for June 2001.

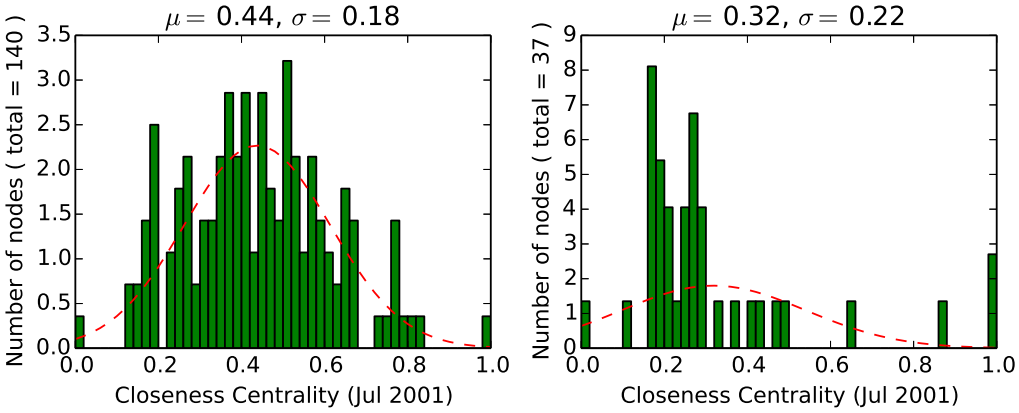


Figure B.19: Plot of Communities' Closeness Distribution for July 2001.

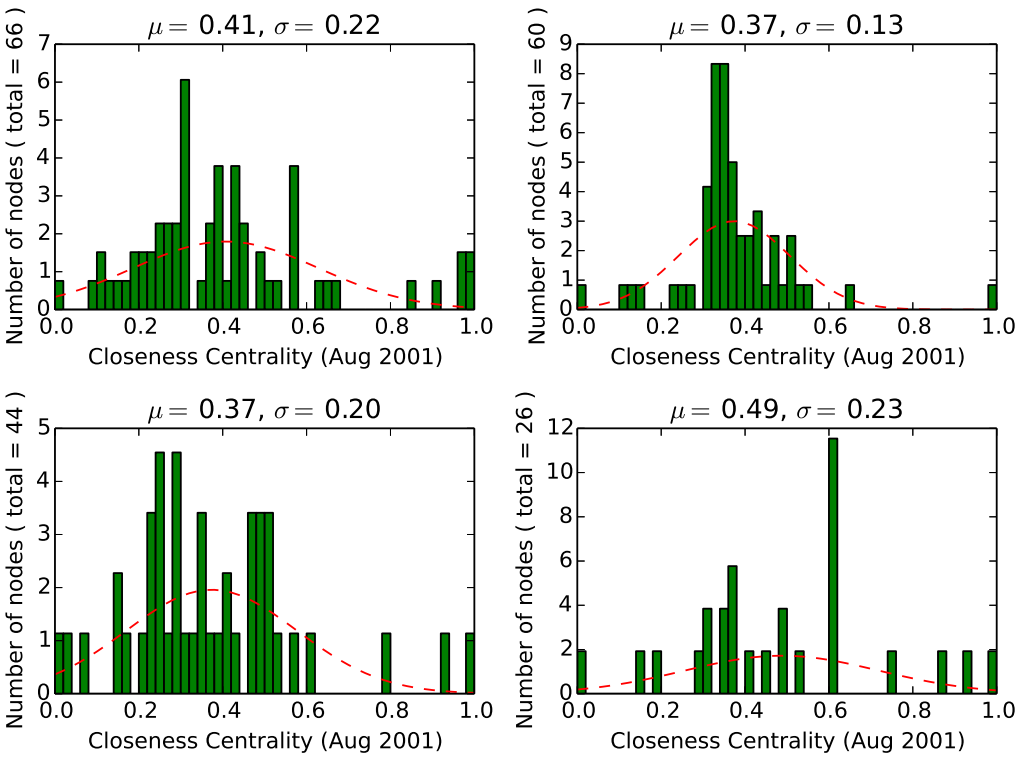


Figure B.20: Plot of Communities' Closeness Distribution for August 2001.

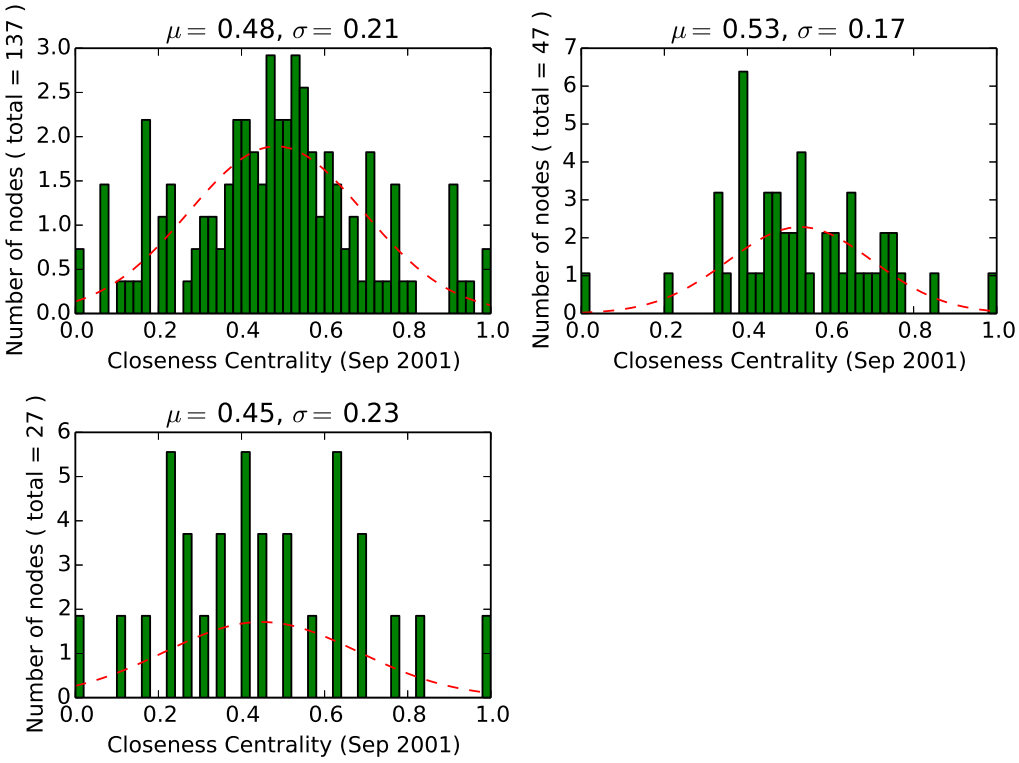


Figure B.21: Plot of Communities' Closeness Distribution for September 2001.

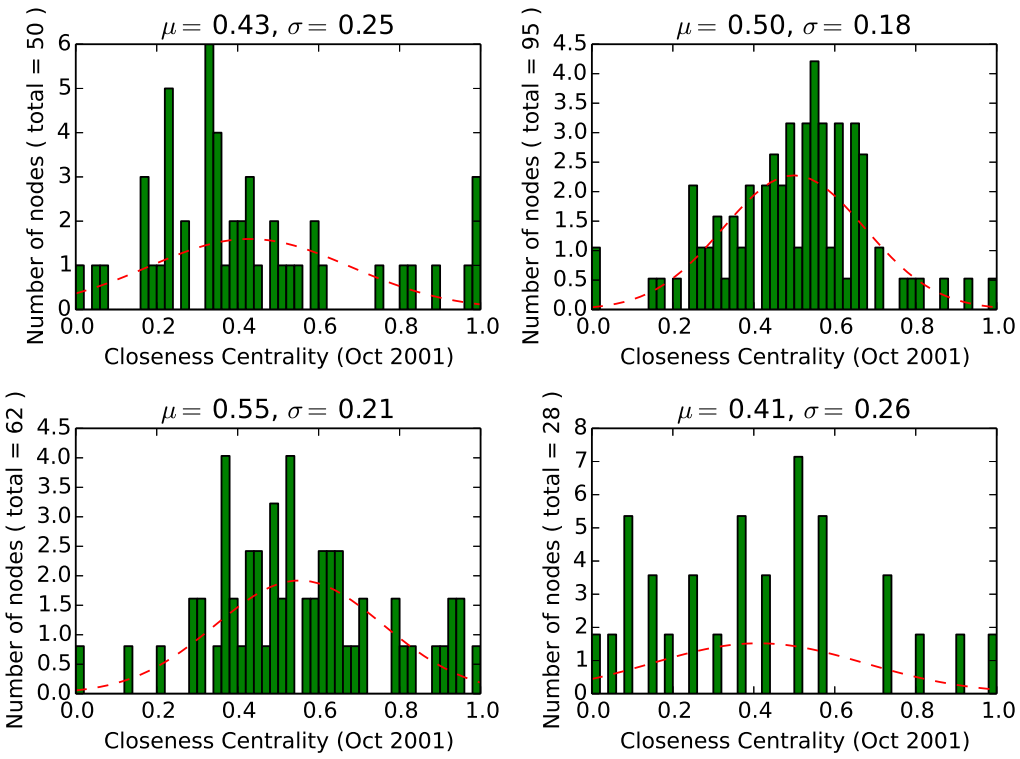


Figure B.22: Plot of Communities' Closeness Distribution for October 2001.

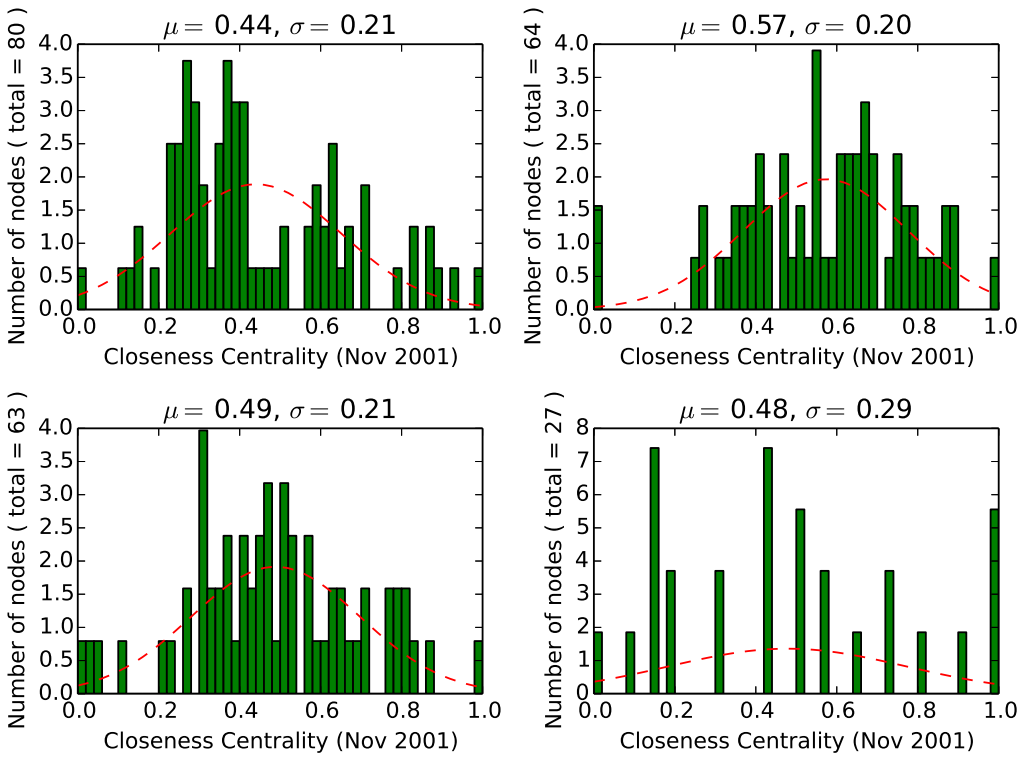


Figure B.23: Plot of Communities' Closeness Distribution for November 2001.

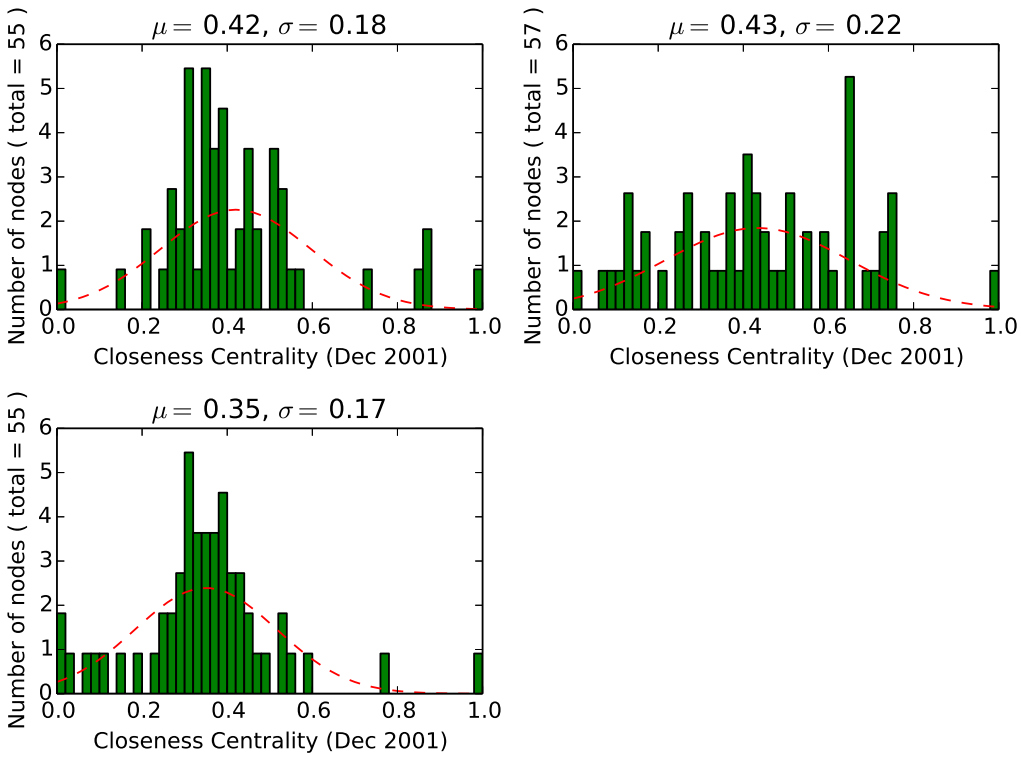


Figure B.24: Plot of Communities' Closeness Distribution for December 2001.

B.3 Mediator Score Distributions

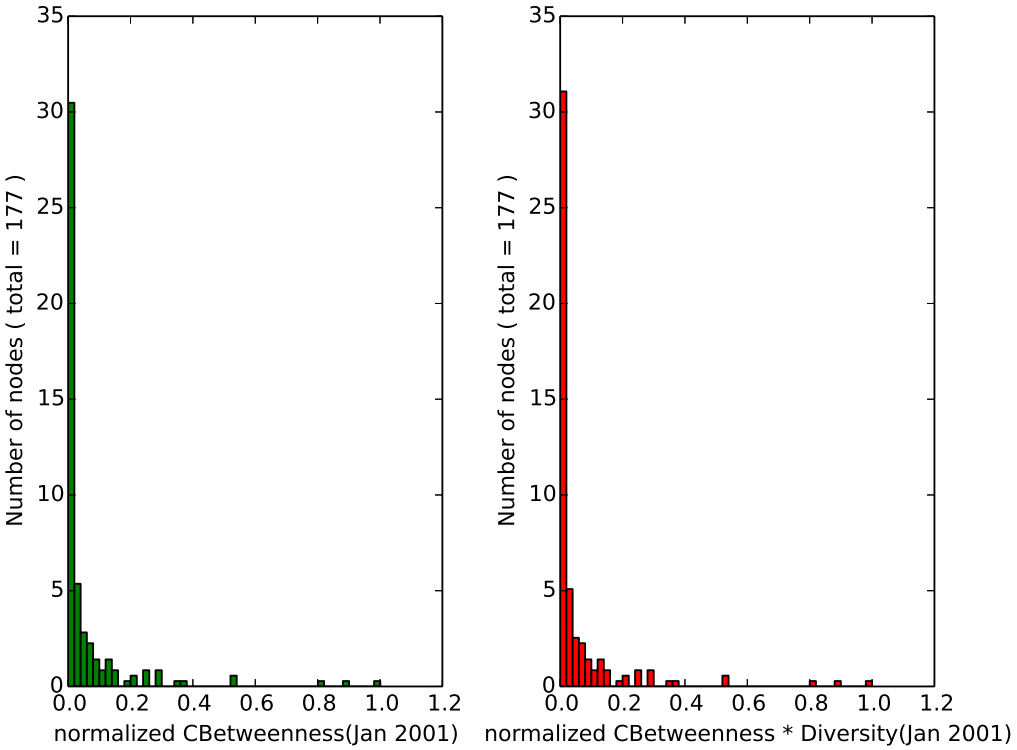


Figure B.25: Plot of Mediator Score Distribution for January 2001.

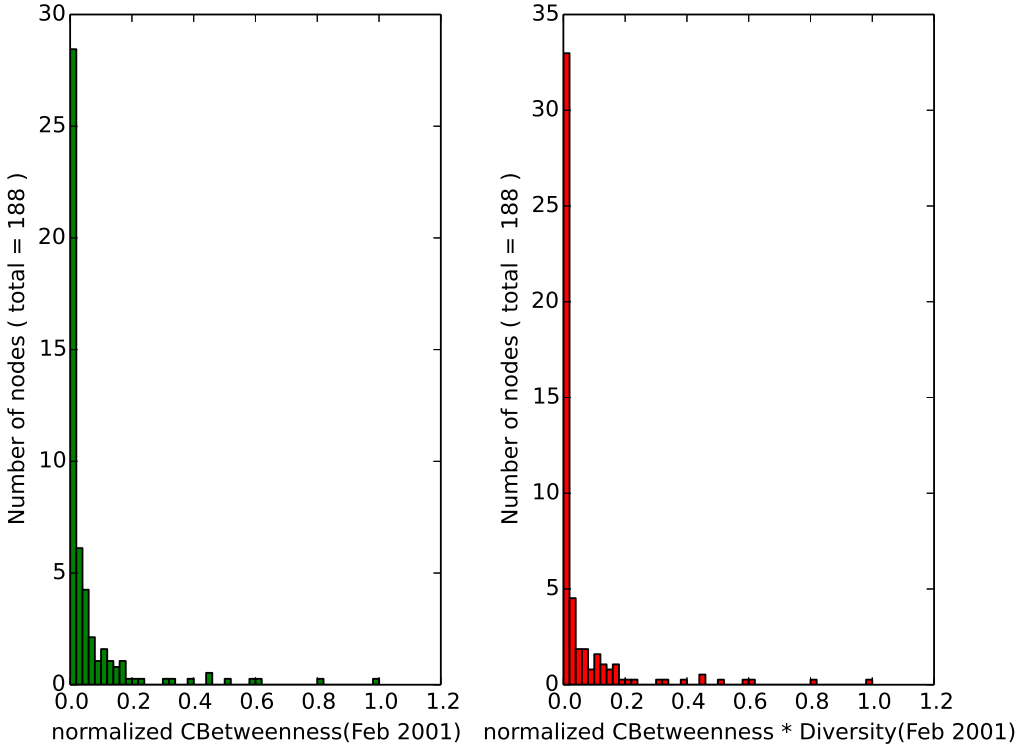


Figure B.26: Plot of Mediator Score Distribution for February 2001.

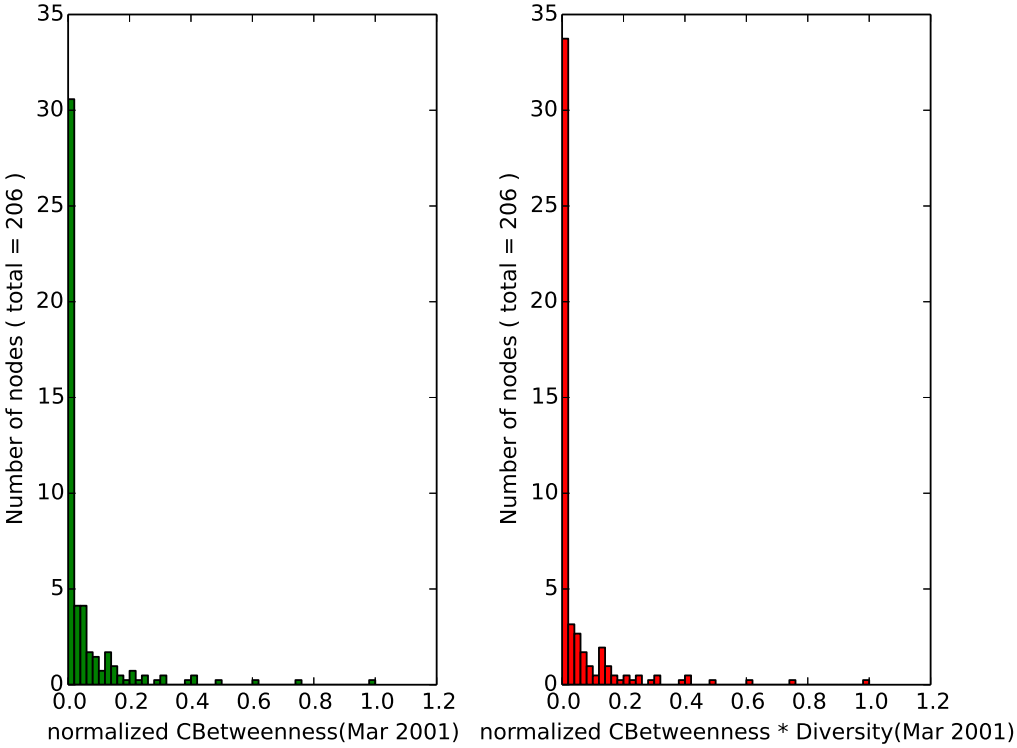


Figure B.27: Plot of Mediator Score Distribution for March 2001.

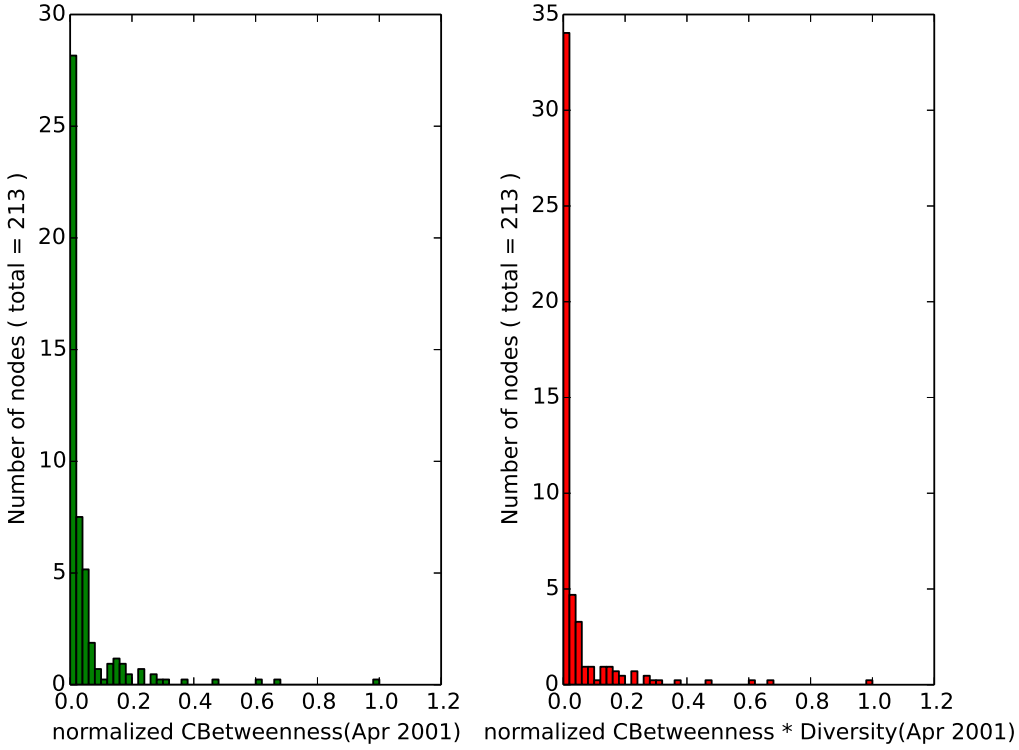


Figure B.28: Plot of Mediator Score Distribution for April 2001.

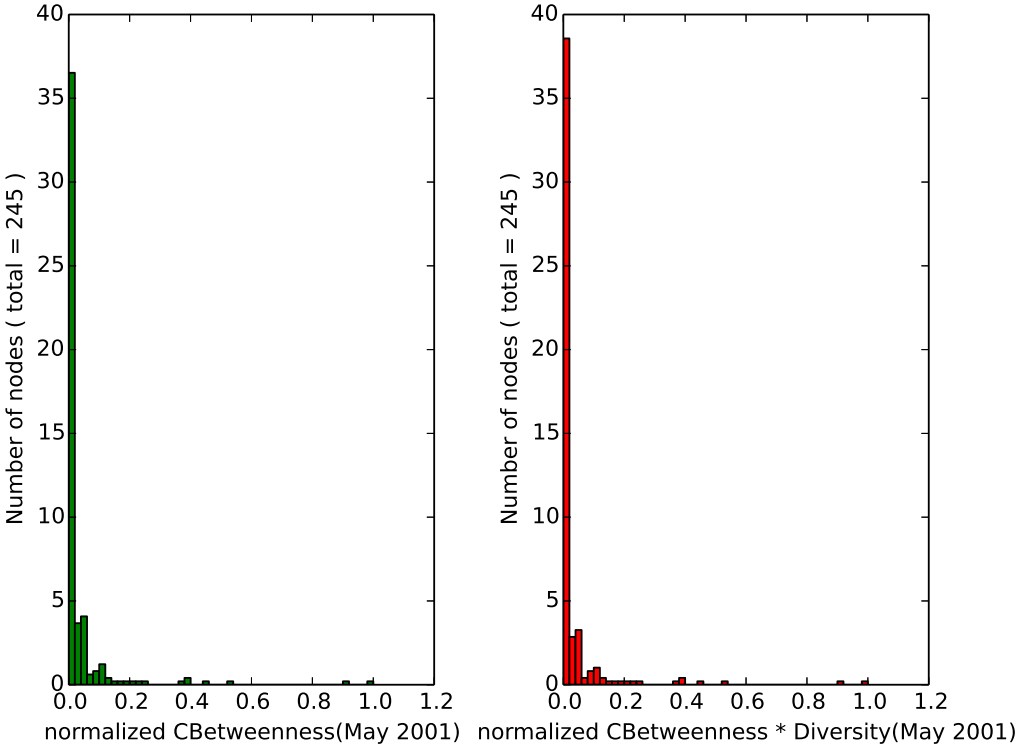


Figure B.29: Plot of Mediator Score Distribution for May 2001.

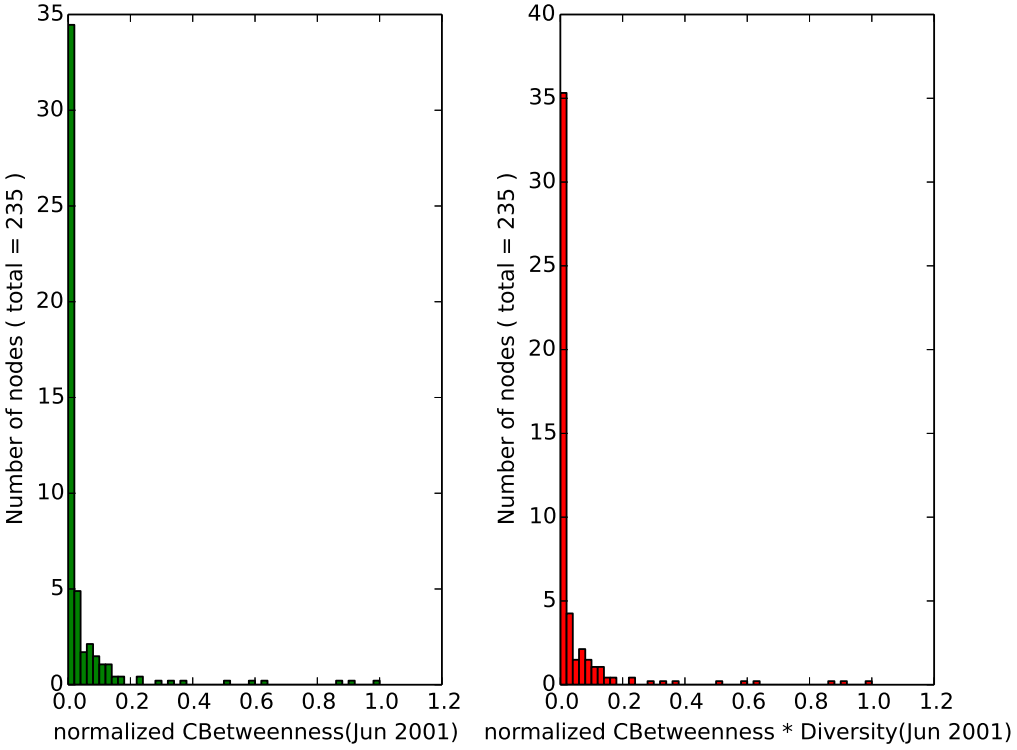


Figure B.30: Plot of Mediator Score Distribution for June 2001.

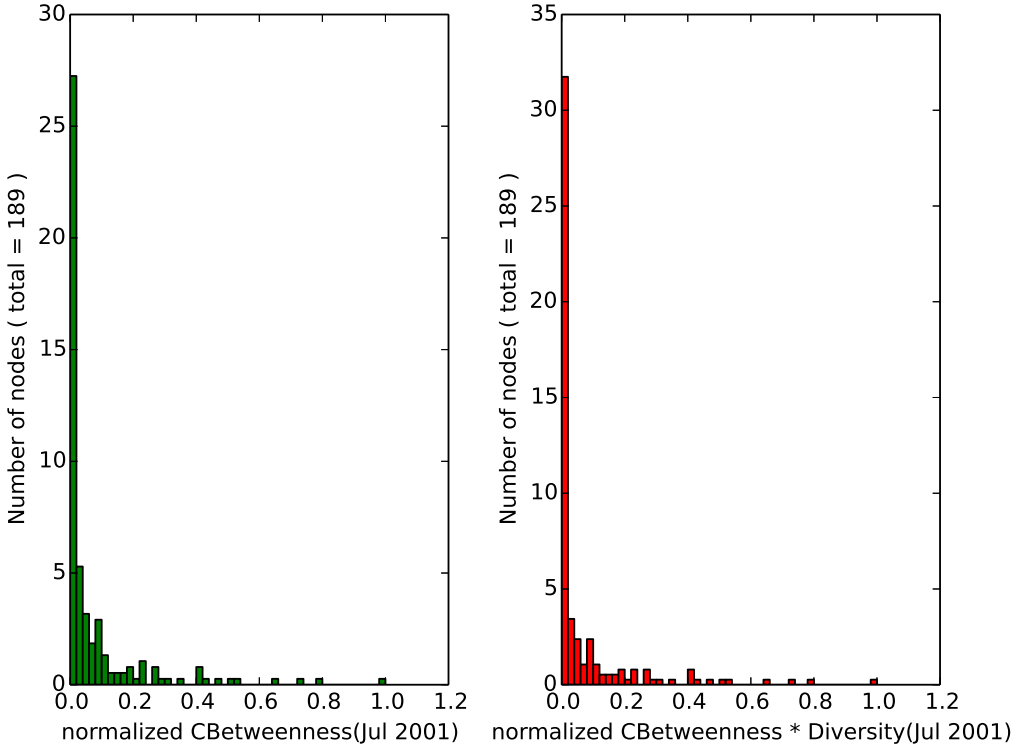


Figure B.31: Plot of Mediator Score Distribution for July 2001.

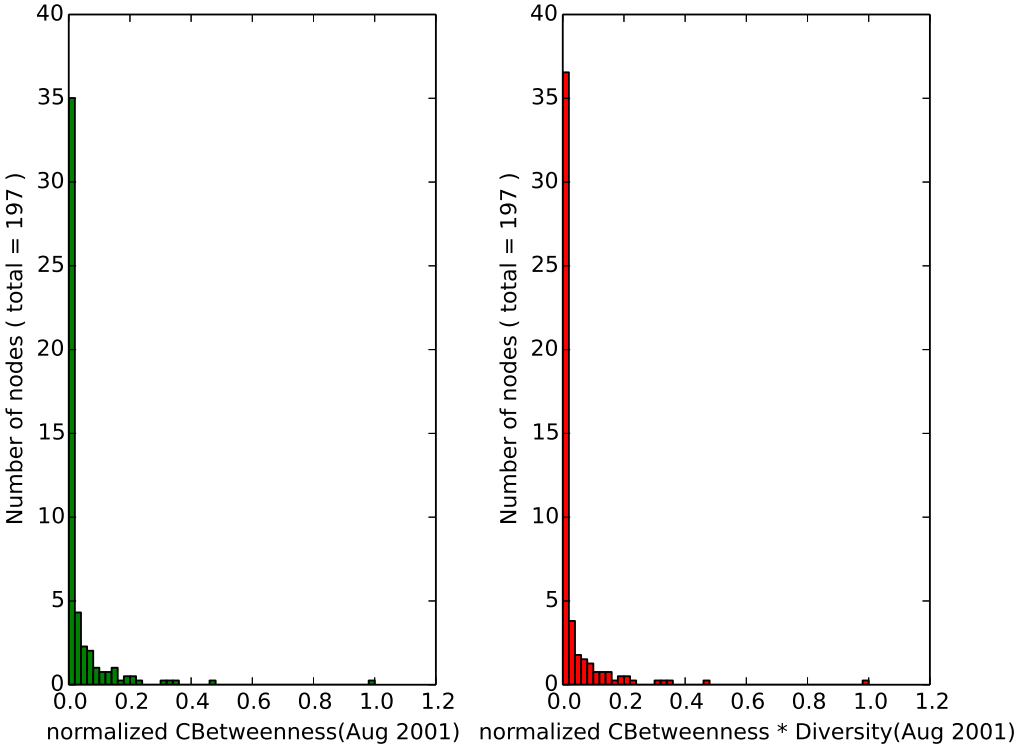


Figure B.32: Plot of Mediator Score Distribution for August 2001.

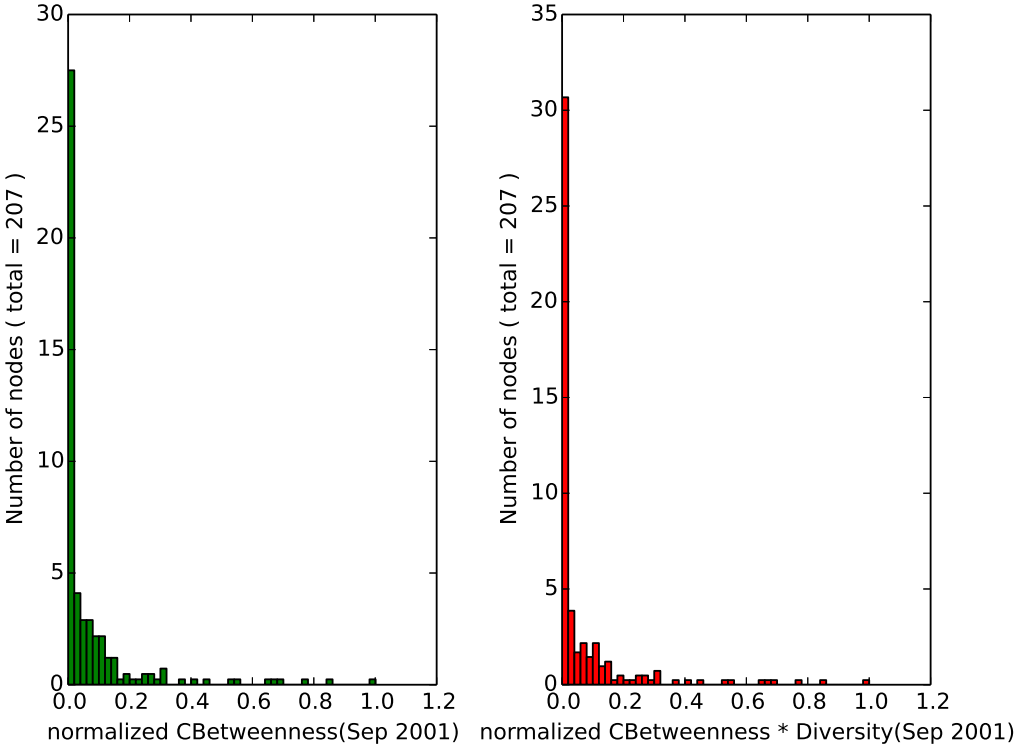


Figure B.33: Plot of Mediator Score Distribution for September 2001.

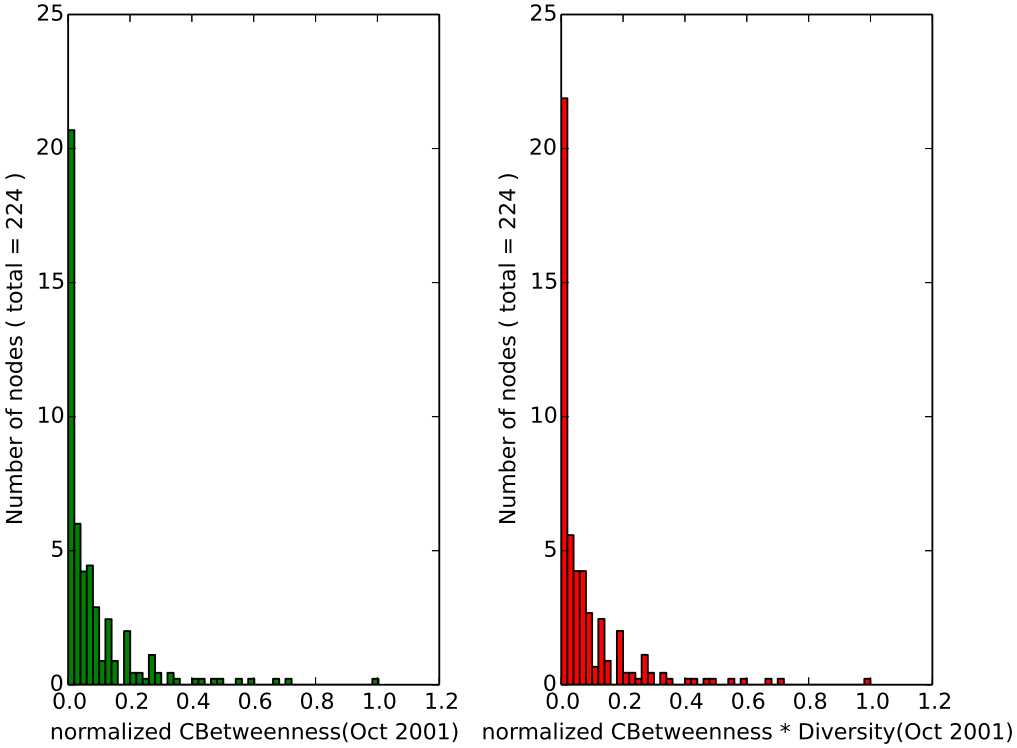


Figure B.34: Plot of Mediator Score Distribution for October 2001.

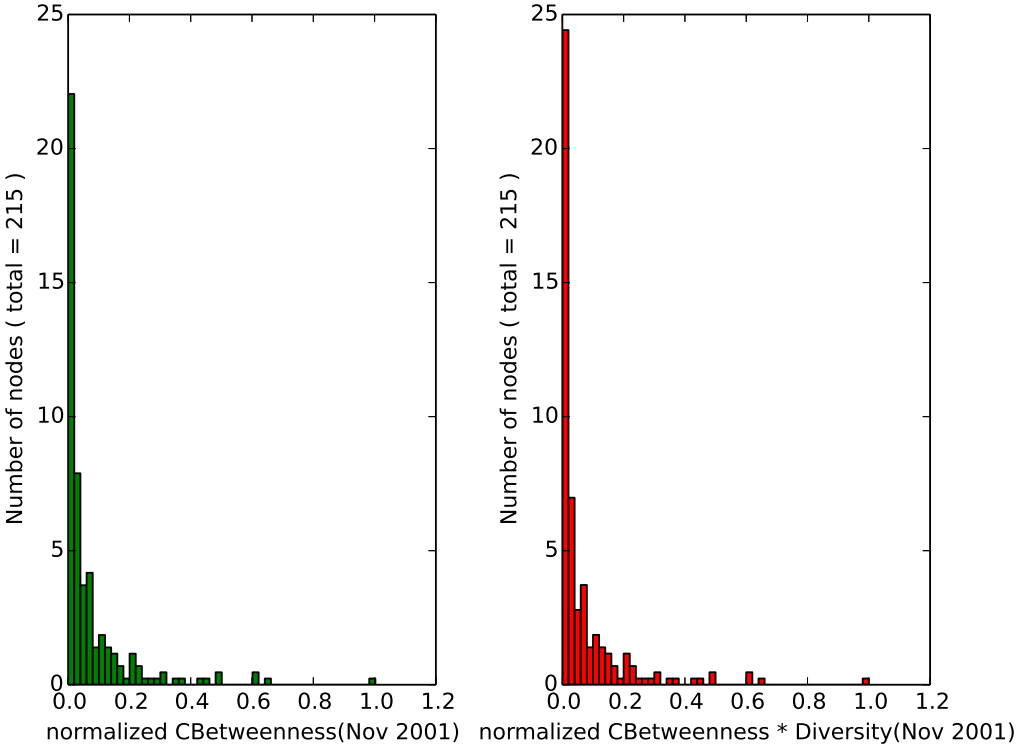


Figure B.35: Plot of Mediator Score Distribution for November 2001.

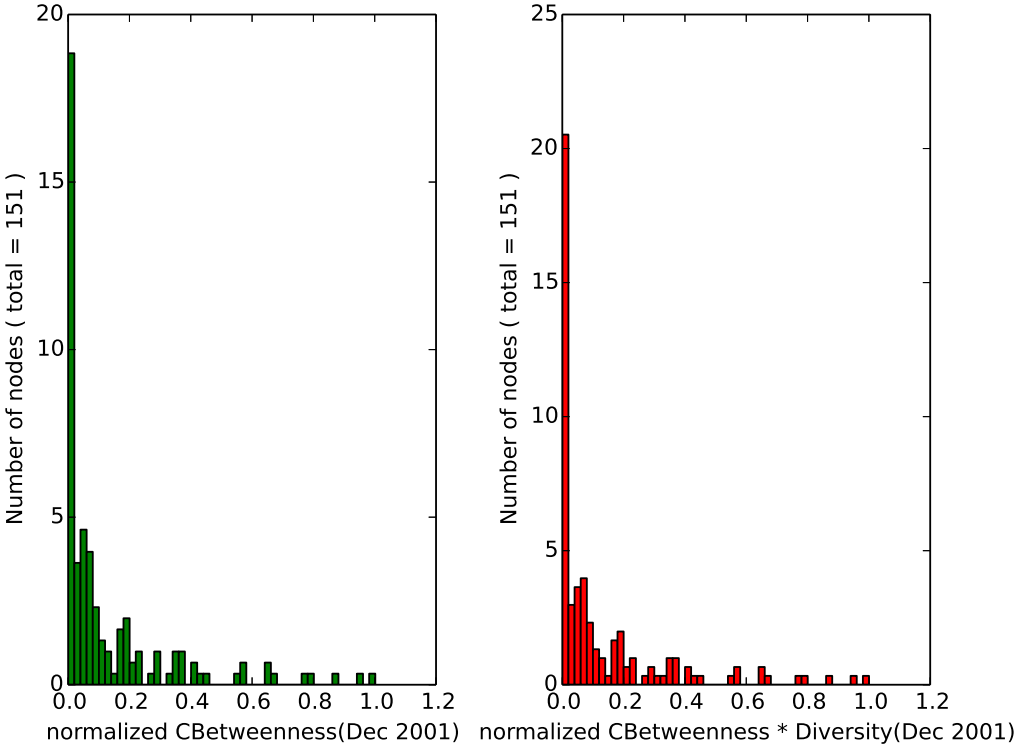


Figure B.36: Plot of Mediator Score Distribution for December 2001.